Technologies des réseaux et NTICs

Le déluge de données
comment en tirer parti

Jeudi 9 juin 2011

Coordination scientifique :

- ***David Menga (EDF R&D)***
- ***Jean-Michel Batto (INRA)***
- ***Pierre Léonard (INRA)***

Amphithéâtre Becquerel, École Polytechnique, Palaiseau

http://www.association-aristote.fr **info@association-aristote.fr**

Edition du 18 prairial an CCXIX (*vulg.* 7 juin 2011)

ARISTOTE Association Loi de 1901. Siège social : CEA-DSI CEN Saclay Bât. 474, 91191 Gif-sur-Yvette Cedex.
Secrétariat : Aristote, École Polytechnique, 91128 Palaiseau Cedex.
Tél. : +33(0)1 69 33 99 66 Fax : +33(0)1 69 33 99 67 Courriel : Marie.Tetard@polytechnique.edu
Site internet http://www.association-aristote.fr

Table des matières

Chapitre 1

Programme de la journée

1.1 Introduction

De nos jours, les données sont partout et constituent la matière première de notre monde numérique. Elles redéfinissent la façon dont on crée de la connaissance scientifique et offrent aux entreprises de nouveaux leviers de croissance et plus de performances opérationnelles aux gouvernements.

Décoder le génome signifie analyser 3 milliards de paires de base. Wallmart traite chaque heure 1 million de transactions clients avec une volumétrie totale de 2,5 Petaoctets. Nous vivons « une révolution industrielle » des données et nos outils actuels de stockage et de traitement des données sont inadaptés pour traiter de tels volumes en un temps acceptable.

Ce séminaire a pour objet de donner les clefs de compréhension de cet univers, de mettre en perspective les défis scientifiques et opérationnels et d'offrir des éléments de réponse à travers des travaux scientifiques de pointe et des solutions marché innovantes.

Le matin sera consacré à la définition de la problématique et donnera des réponses. L'après midi, nous aborderons les expériences des acteurs confrontés à ce défi, comme les électriciens, les biologistes et les astrophysiciens.

1.2 Programme

9h00-9h30	*Accueil café*	
9h30-9h45	**Pierre Léonard** (INRA) **Jean Michel Batto** (INRA) **David Menga** (EDF R&D)	Présentation du séminaire
9h45-10h30	**Xavier Dalloz** (Consultant))	*Big data*, le fuel de l'économie du XXI^{e} siècle
10h30-11h00	*Pause café*	
11h00-11h45	**Vincent Heuschling** (D-FI)	Cartographie des solutions *big data* du marché
11h45-12h30	**Alexandru Costan** (INRIA)	Analyse des systèmes de stockage à grande échelle pour les applications de traitement intensif des données
12h30-14h00	*Repas (salle Detoeuf)*	
14h00-14h35	**Joerg Bienert** (ParStream)	An innovation solution to manage heterogeneous big data
14h35-15h10	**Ronan Keryell** (HPC-Project)	Environnement de programmation pour traitements massifs sur architectures modernes
15h10-15h20	**Denis Caromel** (ActiveEon-INRIA)	Solutions `ProActive` pour *Workflows*, *Map/Reduce*, `Matlab /Scilab`, `CPU/GPU`
15h20-15h45	**Nicolas Pons** (INRA)	La métagénomique, un défi supplémentaire pour la loi de Moore
15h45-16h00	*Pause*	
16h00-16h35	**Patrick Furhmann** (DESY-Hamburg)	`dCache` : scaling out affordable storage.
16h35-17h10	**Marie-Luce Picard** (EDF R&D et ENST-Bilab)	Données massives pour les *smart-grids*
17h10-17h40	**David Konopnicki** (IBM-Haïfa Research Labs)	Massive Scale Analytics for a Smarter Planet
17h45-18h00	**Conclusions, questions-réponses avec les intervenants**	

Chapitre 2

Présentations

Pierre Léonard & Jean Michel Batto (INRA), David Menga (EDF R&D)

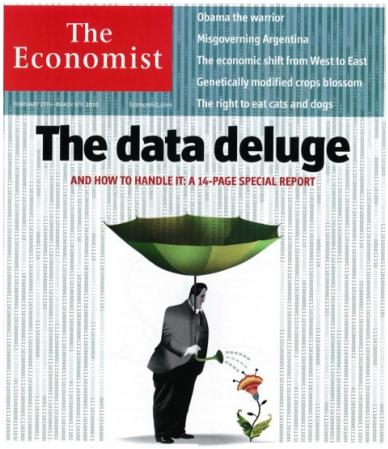
Ouverture du séminaire

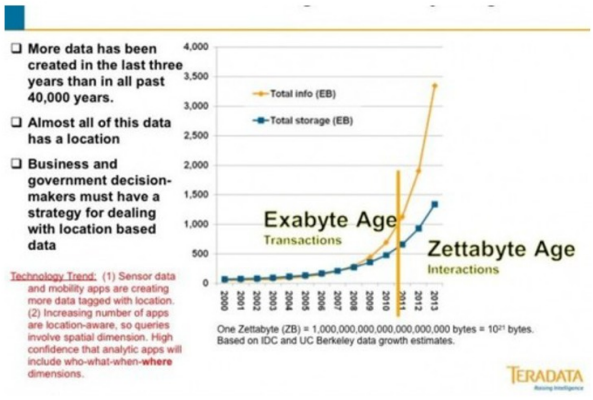
---

aristote

Big Data
Le Déluge de données,
Comment en tirer parti

Jean-Michel Batto, INRA
Pierre Léonard, INRA
David Menga, EDF R&D





Les objets produisent des data en temps réel

- Un seul moteur de Boing produit 10 To d'informations toutes les 30 secondes
- Un voyage de 6 heures New-York/Los Angeles en 737 produit 240 To
- Chaque jour, il y a 28537 vols commerciaux aux Etats Unis

→ 1 Pétaoctet / Jour

Le StorIQ ST32R-1PT , serveur SAN d'Intellique , 1 Petaoctet , coûte 300,000 euros, soit moins de 30 centimes/Go

De nouveaux concepts, BI et BA

- La Business Intelligence ou Informatique décisionnelle désigne les outils et méthodes pour collecter des données et mesurer les performances d'un système à travers des indicateurs bien choisis. Le BI répond aux questions :
  - Que s'est-il passé ?
  - Où sont les problèmes ?
  - Que doit-on faire ?
- Le Business Analytics (BA) fait référence aux technologies et méthodes pour analyser des données afin de comprendre ce qui se passe et de prédire les évolutions futures. Le BA répond aux questions
  - Pourquoi est-ce arrivé ?
  - Que se passera-t-il plus tard si la tendance perdure ?
  - Comment optimiser le système ?

Programme

09h00-09h30	**Accueil café**	
9h30-9h45	**Pierre Léonard, Jean Michel Batto (INRA), David Menga** (EDF R&D)	Présentation du Séminaire
9h45-10h30	**Xavier Dalloz (Consulting)**	*Big data*, le fuel de l'Economie du XXI éme siècle
10h30-11h00	**Pause café**	
11h00-11h45	**Vincent Heuschling (D-FI)**	Cartographie des solutions *big data* du marché
11h45-12h30	**Alexandru Costan (doctorant INRIA)**	Quel système de gestion de données pour stocker les données massives ?
12h30-14h00	**Déjeuner (salle Detoeuf)**	

14h00-14h35	**Joerg Bienert (Parstream)**	An innovation solution to manage heterogeneous big data
14h35-15h10	**Emmanuel Chiva (HPC-Project)**	Quel environnement de programmation pour traiter les big data
15h10 -15h 20	**Denis Caromel (ActiveEon-INRIA)**	Solutions ProActive pour Workflows MapReduce, Matlab/Scilab, CPU/GPU
15h 20-15h45	**Nicolas Pons (INRA)**	La mégagénomique, un défi supplémentaire pour la loi de Moore
15h 45- 16h 00	**Pause**	
16h 00-16h35	**Patrick Furhmann (/DESY)**	dCache : scaling out affordable storage
16h35-17h10	**Marie-Luce Picard (EDF R&D et ENST-Bilab)**	Données massives pour les Smart-grids
17h 10- 17h45	**David Konopnicki (IBM Haïfa-Research Lab)**	Massive Scale Analytics for a Smarter Planet
17h45-18h00	**Conclusion, questions réponses avec les intervenants**	

Le Doomsday Book (1086)

- Le **Domesday Book** (ou simplement **Domesday**), en français *Livre du Jugement Dernier*[1], est l'enregistrement du grand inventaire de l'Angleterre terminé en 1086, réalisé pour Guillaume le Conquérant, l'équivalent de nos jours d'un recensement national.

2.1 Xavier DALLOZ (Consultant)

Big data, le fuel de l'économie du XXI^{e} siècle

Régulièrement une innovation majeure change tout avec à chaque fois de nouvelles technologies, de nouveaux métiers et de nouveaux enjeux. Après l'ère des *mainframes*, l'ère des *mini computers*, l'ère des PC, l'ère du *software* pour améliorer la productivité personnelle, l'ère de l'Internet, voici celle du « *Big Data* ».

Le *Big Data* va permettre notamment de repenser les modèles économiques en misant sur de nouvelles créations de valeur avec notamment l'intelligence collaborative et la *shazamisation* de notre environnement de telle sorte qu'il y ait une meilleure efficacité de « notre » capital : santé, énergie, éducation, équipements, stocks... Avec le *Big Data*, la chasse aux gaspillages va enfin devenir une réalité. Nous n'avons encore rien vu... Tout va s'accélérer.

Que d'opportunités ! Que de leviers de croissance pour nos économies !

Les datas = le fuel du 21ième sicècle

D'énormes gisements de création de valeurs

http://www.yournetworkmarketing.com/facebook-twitter-youtube-stats-in-real-time-simulation/

Xavier Dalloz

dalloz@dalloz.com

XAVIER DALLOZ CONSULTING

Le Plan

- **Définition des Big Data : un réducteur de la complexité + une aide à la prise de la décision et au management du risque (intuition)**
- **Exemples de Big Data**
- **Les outils du Big Data**
- **L'enjeu : la création de la valeur en optimisant l'utilisation du capital fixe, circulant et immatériel**
- **Exemples d'usage des Big Data**
  - Le commerce de détail (stocks)
  - La production des biens manufacturés (co-création)
  - La géolocalisation (traces)
  - L'Internet des objets et la shazamisation (conversation)
  - Le web sémantique (meta data)
- **Recommandations**

dalloz@dalloz.com

XAVIER DALLOZ CONSULTING

Définition des Big Data (Wikipedia)

Une définition qui change avec le temps...

- Ensembles de données qui deviennent tellement gros qu'ils en deviennent difficiles à travailler avec des outils classiques de gestion de base de données.
- **Les perspectives du traitement des big data sont énormes**, notamment pour l'analyse d'opinions ou de tendances industrielles, la génomique, l'épidémiologie ou la lutte contre la criminalité.
- La production de données par les utilisateurs, et notamment le **partage d'informations ubiquitaires** (capteurs et senseurs mobiles, caméras, microphones, appareils photos, lecteurs RFID, réseaux de capteurs sans fil, etc.) augmentent drastiquement le nombre de données pouvant être traitées

dalloz@dalloz.com

XAVIER DALLOZ CONSULTING

Une autre définition

- Le Big Data est le moyen de valider une intuition
- En validant une intuition le Big Data est un réducteur de risques.
- Pour cela, il faut traiter un déluge de données en temps réel.
- Le grand frein du Big Data est l'incapacité à définir clairement notre intuition. Il faut être problématicien.

dalloz@dalloz.com

XAVIER DALLOZ CONSULTING

Ce déluge de données a de nouvelles propriétés

Plus de données ont été créé ces 3 dernières années que pendant les 40.000 années précédentes

- Ce sont des données **non structurées**
- Elles sont produites en **temps réel**
- Elles arrivent mondialement en **flots continus**
- Elles sont **méta taguées** mais de façon disparate (localisation, heure, jour, etc.)
- Elles proviennent de sources très **disparates** (téléphone mobile, capteurs, téléviseurs connectés, tablettes, PC fixes, PC portables, objets, machines), de façon **désordonnée** et **non prédictible**.

dalloz@dalloz.com

XAVIER DALLOZ CONSULTING

Unit	Size	What it means
Bit (b)	1 or 0	Short for "binary digit", after the binary code (1 or 0) computers use to store and process data
Byte (B)	8 bits	Enough information to create an English letter or number in computer code. It is the basic unit of computing
Kilobyte (KB)	1,000, or 2^{10}, bytes	From "thousand" in Greek. One page of typed text is 2KB
Megabyte (MB)	1,000KB; 2^{20} bytes	From "large" in Greek. The complete works of Shakespeare total 5MB. A typical pop song is about 4MB
Gigabyte (GB)	1,000MB; 2^{30} bytes	From "giant" in Greek. A two-hour film can be compressed into 1-2GB
Terabyte (TB)	1,000GB; 2^{40} bytes	From "monster" in Greek. All the catalogued books in America's Library of Congress total 15TB
Petabyte (PB)	1,000TB; 2^{50} bytes	All letters delivered by America's postal service this year will amount to around 5PB. Google processes around 1PB every hour
Exabyte (EB)	1,000PB; 2^{60} bytes	Equivalent to 10 billion copies of *The Economist*
Zettabyte (ZB)	1,000EB; 2^{70} bytes	The total amount of information in existence this year is forecast to be around 1.2ZB
Yottabyte (YB)	1,000ZB; 2^{80} bytes	Currently too big to imagine

Source: *The Economist*

The prefixes are set by an intergovernmental group, the International Bureau of Weights and Measures. Yotta and Zetta were added in 1991; terms for larger amounts have yet to be established.

dalloz@dalloz.com

XAVIER DALLOZ CONSULTING

	What was measured	Amount of data	Year of estimate
MGI storage-based approach	• New data stored in enterprise external disk storage in a year • New data stored by consumers in a year	• 7.4 x 10^{18} bytes (includes replicas) • 6.8 x 10^{18} bytes	• For 2010
IDC/EMC[1] Digital Universe	• Annual digital data captured (includes all generated, stored or not) • Includes more than 60 types of devices • Did not include information consumption by users through TV, video gaming[1]	• ~800 x 10^{18} bytes	• For 2009
UCSD	• Includes both digital and analog data for TV, radio, phone, print, computer, comp. games, movies, recorded music, etc. • Measured data from consumption perspective[2]	• 3.6 x 10^{21} bytes (total consumption US only)	• For 2008
Hilbert, López	• Capacities for specific technologies – Server and mainframe hard disks – Other hard disks – Digital tape – PC hard disk • Total digital storage capacity	• 24.5 x 10^{18} bytes • 6.49 x 10^{18} bytes • 32.5 x 10^{18} bytes • 123 x 10^{18} bytes • 276 x 10^{18} bytes	• For 2007

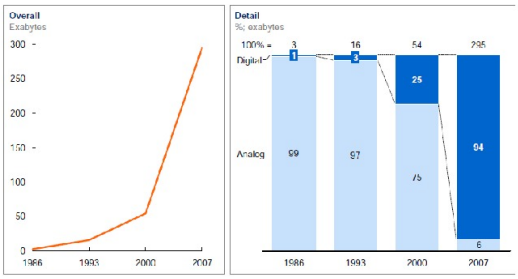
1 Includes chip cards, floppy disks, camera, video games, mobiles, memory cards, media players, CDs, DVDs, Blue Ray disks, PC and server hard disks.
2 Consumption is defined as the data each time used by the user.
SOURCE: IDC white papers on Digital Universe, sponsored by EMC; Bohn and Short, *How Much Information? 2009: Report on American Consumers*, January 2010; Hilbert and López, "The world's technological capacity to store, communicate, and compute information," *Science*, February 2011; McKinsey Global Institute analysis

dalloz@dalloz.com

XAVIER DALLOZ CONSULTING

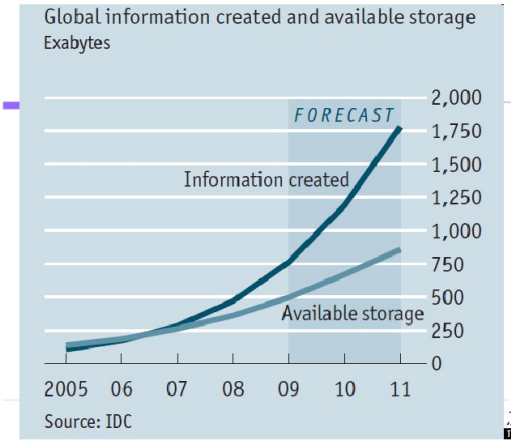
Data storage has grown significantly, shifting markedly from analog to digital after 2000

Global installed, optimally compressed, storage

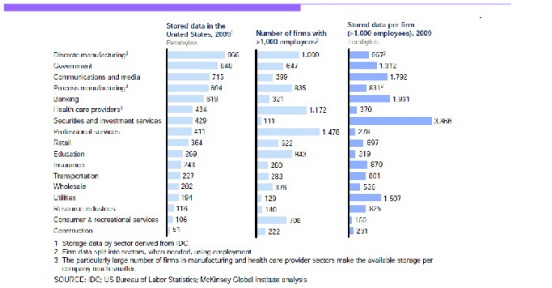


dalloz@dalloz.com

XAVIER DALLOZ CONSULTING



Le tas d'or sur lequel les entreprises sont assises...

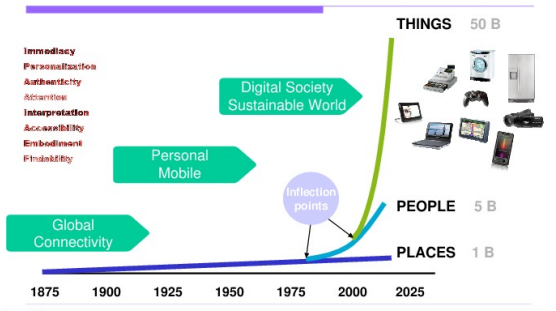


dalloz@dalloz.com

XAVIER DALLOZ CONSULTING

50 milliards d'objets connectés en 2025

des nouvelles créations de valeur

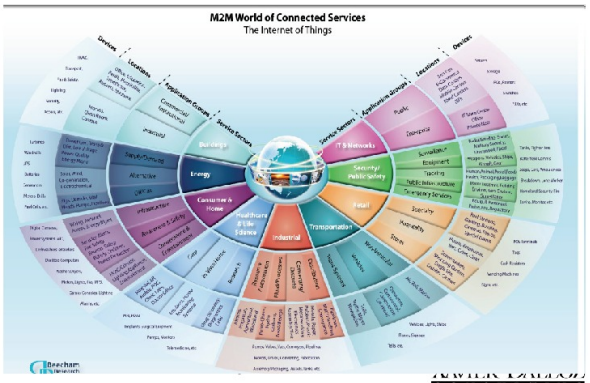


Source: Ericsson

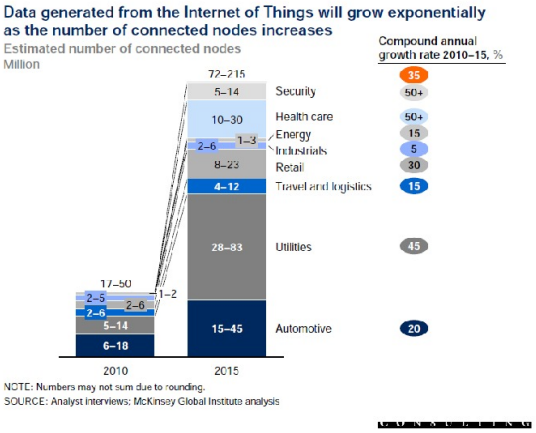
dalloz@dalloz.com

XAVIER DALLOZ CONSULTING

Nous n'avons encore rien vu !!



XAVIER DALLOZ CONSULTING



L'interaction avec notre environnement

Vers une économie de l'immatériel et de l'intelligence collaborative

« Le job en vogue dans les 10 prochaines années sera statisticien… La faculté de prendre des données— et être capable de les comprendre, de les traiter, d'en extraire de l'information, ainsi que de les visualiser et de les communiquer »

Hal Varian

(Google's chief economist)

Nouvelles technologies, nouveaux métiers, nouveaux enjeux…

- **De nouveaux enjeux**
  - Comment exploiter ces nouveaux volumes de données ?
  - Comment les stocker?
  - Comment les traiter?
  - Comment les visualiser?
- **Et de nouvelles technologies**
  - Bases de données distribuées
  - Traitement de données distribué
  - Analyse d'événements en temps réel
  - "Cloud Computing"
  - Bases de données distribuées
  - Traitement de données distribué
  - Analyse d'événements en temps réel
  - "Cloud Computing"

Exemples d'outils

- A/B testing
- Association rule learning
- Classification
- Cluster analysis
- Crowdsourcing
- Dara fusion and data integration
- Data mining
- Ensemble learning
- Genetic algorithms
- Machine learning
- Natural language processing
- Neural Networks
- Network analysis
- Pattern recognition
- Predective modeling
- Regression
- Sentiment analysis
- Signal processing
- Spatial analysis
- Statistics
- Supervised learning
- Simulation
- Time series analysis
- Unsupervised learning
- Visualization

Le plus de données ouvertes, le plus d'intelligence à construire…

- Big Table
- Business intelligence
- Cassandra
- Data mart
- Data warehouse
- Hadoop
- MapReduce
- Metadata
- R.

L'enjeu : créer de la valeur

- **Il faut exploiter les Big Data** non pour automatiser le passé,
- Mais pour faire:
  - **Plus** (productivité, pénétration…)
  - **Mieux** (compétitivité, pertinence…)
  - **Autre chose** (nouveaux marchés, innovations, partenariats, offres contextuelles…)
  - **Différemment**

dalloz@dalloz.com

XAVIER DALLOZ CONSULTING

L'or des poussières

- **Les techniques du Big Data facilitent des interactions**
  - Moins chères
  - Plus rapides
  - Plus efficaces
  - Moins contraintes par le temps et les distances.
  - D'où de nouveaux modèles économiques
- **De plus en plus de stratégies gagnantes sont basées sur la valorisation**
  - de la poussière d'information (Google),
  - de petites transactions (eBay),
  - de petites publicités (Google encore),
  - de collaborations (logiciels libres)…
  - Appliquer à l'envers la règle du 80/20 pour transformer la poussière en or…

dalloz@dalloz.com

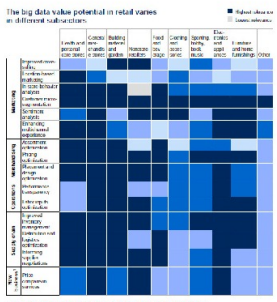
XAVIER DALLOZ CONSULTING

Exemples d'usage des Big Data

dalloz@dalloz.com

XAVIER DALLOZ CONSULTING

Toutes les secteurs d'activité économique sont concernées…



dalloz@dalloz.com

XAVIER DALLOZ CONSULTING

Les nouveaux acteurs utilisent massivement les Big Data

- **Zynga**
  - Créateur des jeux sociaux les plus populaires 10% de la population Internet mondiale a joue un jeu Zynga (230M joueurs / mois) Déplace 1Pb de données chaque jour Ajoute 1000 serveurs par semaine
- **Facebook**
  - 500M d'utilisateurs
  - 3.5B morceaux de contenu / semaine
  - 4B messages / jour
  - 1.2M photos / second (lues)
- **Twitter**
  - 70M Tweet / day
    = 800 tweets/sec * 200b = 160kb/sec = 9Mb/min = 12Gb/jour de texte 8T de données / day 6B api calls / day

dalloz@dalloz.com

XAVIER DALLOZ CONSULTING

Le commerce de détail

- La base de données de WalMart contient 2,5 petabytes :
  - CPFR
  - EDLP
  - Retail Link
- TESCO collecte 1,5 milliards de data chaque mois pour ajuster ses prix et les promotions
- eBay a une base de données de 6,5 et un datawarehouse de 2,5 petabyte

dalloz@dalloz.com

XAVIER DALLOZ CONSULTING

Big data retail levers can be grouped by function

Function	Big data lever
Marketing	▪ Cross-selling ▪ Location based marketing ▪ In-store behavior analysis ▪ Customer micro-segmentation ▪ Sentiment analysis ▪ Enhancing the multichannel consumer experience
Merchandising	▪ Assortment optimization ▪ Pricing optimization ▪ Placement and design optimization
Operations	▪ Performance transparency ▪ Labor inputs optimization
Supply chain	▪ Inventory management ▪ Distribution and logistics optimization ▪ Informing supplier negotiations
New business models	▪ Price comparison services ▪ Web-based markets

SOURCE: McKinsey Global Institute analysis

CONSULTING

La production des biens manufacturés (vendre avant de produire)

dalloz@dalloz.com

XAVIER DALLOZ CONSULTING

L'indicateur de base

$$\frac{\text{Résultat d'exploitation}}{\text{Capitaux fixe + circulant}} = \frac{\text{Résultat d'exploitation}}{\text{Chiffre d'affaire}} \times \frac{\text{Chiffre d'affaire}}{\text{Capitaux fixe + circulant}}$$

Rentabilité économique
Elle croît si:
- ❑Qualité, différence de l'offre,
- ❑Efficacité de la relation client,
- ❑Offre nouvelle,
- ❑Nouvelles clientèles, extension territoriale

Augmente aussi si:
Efficacité accrue de l'utilisation du capital

Taux de marge
Il croît si l'on augmente
- ❑ la rentabilité des ventes
- ❑ leur pertinence (sur mesure)
- ❑ leur différentiation
- ❑ Leur valeur ajoutée

Taux de rotation du capital d'exploitation
la rentabilité financière croît avec
- ❑la vitesse de rotation des actifs
- ❑La réduction du capital fixe nécessaire (partenariats, entreprise étendue…)

dalloz@dalloz.com

XAVIER DALLOZ CONSULTING

3 impératifs liés par les Big Data: gérer, produire, vendre

- Gérer: mieux exploiter le capital
  - Le capital fixe
  - Le capital circulant
  - L'intelligence collective interne
  - Les synergies de réseau à valeur ajoutée
  - Les risques et opportunités (innovations, acquisitions…)
- Produire: réduire les coûts
  - Des matériaux, composants, équipements,
  - Personnels, management, frais généraux…
  - Coûts d'interactions
- Vendre: accroître les revenus
  - Parts de marché, nouveaux marchés
  - Clients:fidélisation, acquisition et extension
  - Time to market, cycle de renouvellement

dalloz@dalloz.com

XAVIER DALLOZ CONSULTING

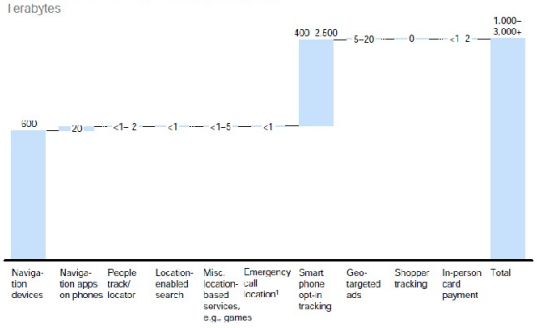
La géolocalisation

dalloz@dalloz.com

XAVIER DALLOZ CONSULTING

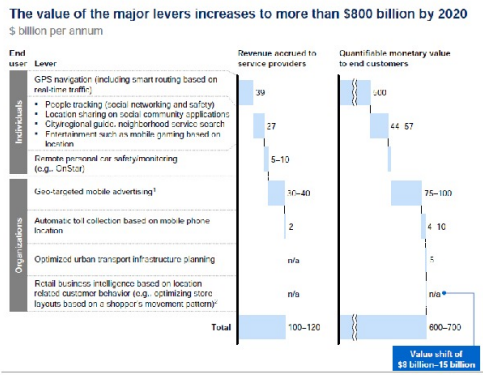
Overall personal location data total more than 1 petabyte per year globally

Personal location data generated globally, 2009
Terabytes



dalloz@dalloz.com

XAVIER DALLOZ CONSULTING



dalloz@dalloz.com

XAVIER DALLOZ CONSULTING

L'Internet des objets et la shazamisation

- Tags RFID
  - 12 millions vendus en 2011
  - 209 milliards en 2021
- De nouveaux process métiers (traçabilité)
  - Computer aided engineering
  - Computer aides manufacturing
  - Collaborative product development management
  - Digital manufacturing
  - Etc.

dalloz@dalloz.com

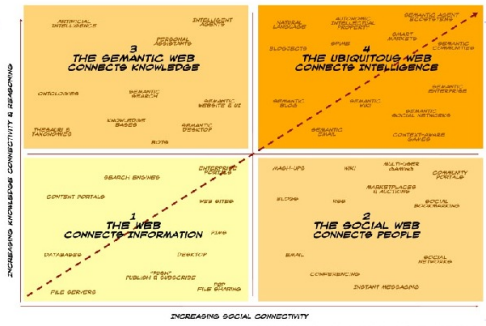
XAVIER DALLOZ CONSULTING

Le web sémantique

dalloz@dalloz.com

XAVIER DALLOZ CONSULTING

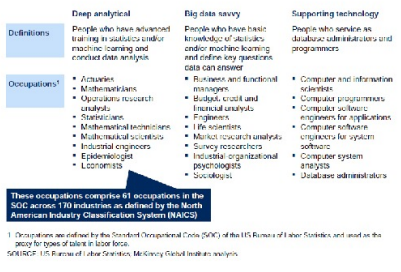
What is the evolution of the internet to 2020?



dalloz@dalloz.com

XAVIER DALLOZ CONSULTING

Recommandations



dalloz@dalloz.com

XAVIER DALLOZ CONSULTING

Devenir problématicien…

- Il faut apprendre à poser la bonne question
- Trop d'informations… tuent l'information
- Réduire la complexité
- Mesurer pour valider les intuitions
- Les principaux gisements de création de valeurs sont :
  - "Predictive analysis"
  - Rendre l'information intelligible
  - Visualisation de données

dalloz@dalloz.com

XAVIER DALLOZ CONSULTING

2.2 Vincent HEUSCHLING (D-FI)

Cartographie des solutions *big data* du marché

Face au déluge de donnée, que nous vivons aujourd'hui, quelles sont les réponses des grands acteurs du marché ? Quels défis présente cette explosion du volume de données pour les infrastructures. Décryptage des offres des grands constructeurs que sont `EMC`, `IBM`, `Oracle`, ..., et des architectures innovantes du monde opensource.

Cartographie des solutions BigData

Panorama du marché
et prospective

Solutions BigData

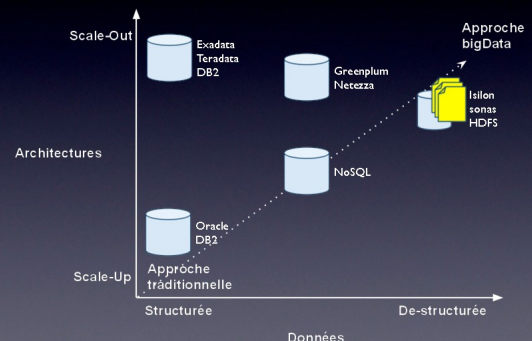
- **Défi(s) pour les fournisseurs**
- **Quel marché**
- **Architectures**
- **Acteurs commerciaux**
- **Solutions alternatives**

Quels Défis ?

- **des volumes impossibles à traiter :**
  - 30 To de logs par jour chez Facebook
  - 15 Po de data par an au CERN
- **des croissances vertigineuses**
- **du business en temps réel**
- **des données différentes :**
  - Non structurées, réparties , NoSQL...

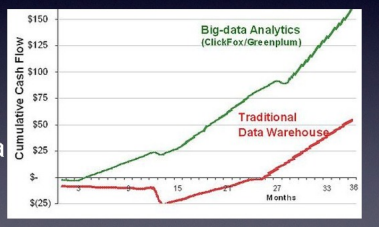
Le quadrant magique
(DW database management)

Positionnement des acteurs du marché



ROI

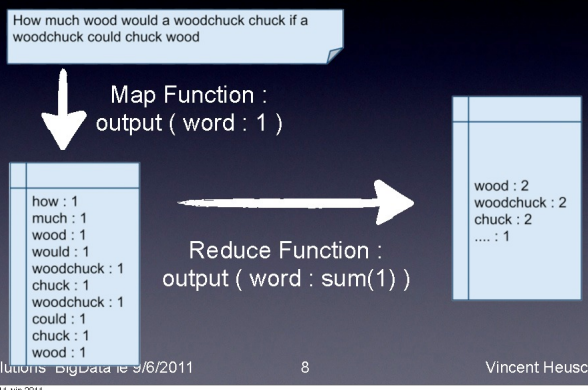
- Révolutionne les datawarehouses existants
- ROI de 27 mois à 6 mois
- 3 fois moins cher
- 4 fois plus rapide à implémenter



Architecture & composants

- Shared Disk vs Share Nothing Arch.
- Hadoop / HBase / HDFS
- Map Reduce

Map Reduce



Map Reduce

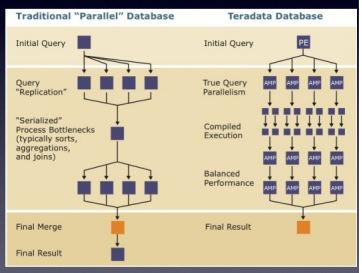
- S'appuie sur une base key / value
- est scalable sur n serveurs
- permet d'enchainer plusieurs Reduce
- beaucoup d'implémentations

Acteurs du marché

- **Teradata**
- **Oracle / Exadata**
- **IBM / Netezza**
- **EMC / Greenplum**
- **...**

Teradata

- Depuis 1979
- Appliances
- Share nothing arch.
- Parallélisme
- Pour les DW
- De 6 To à 92 Po

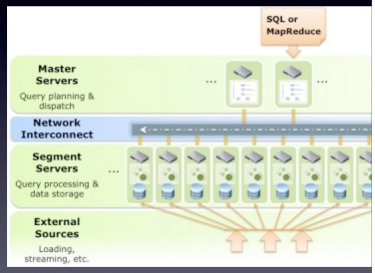


Oracle Exadata

- «Database machine» (n'est pas une appliance)
- Serveurs de stockage (168 cores, 5 TB de flash cache, 45 TB utiles)
- Serveurs de traitements (128 cores / 2 TB de mémoire)
- 1500000 IOPS
- Data Load Rate: Up to 12 TB/hour

EMC Greenplum

- Serveurs std
- Share nothing arch
- Map Reduce
- SQL



IBM Netezza

- Blades IBM + Disques + FPGAs
- Share nothing arch.
- Map Reduce & SQL
- Data load rates de 2TB/h
- Produits : Skimmer (1TB à 10TB) & TwinFin (1TB à 1PB+)

Alternatives et Opensource

- **Active circle**
- **Bases NOSQL**
- **Apache HADOOP**
- **Database.com**
- **Amazon Elastic Map Reduce**

Active Circle

- FileSystem distribué
- Accès par NAS ou API
- Virtualisation sur disque et bande
- Noeuds locaux ou distants
- Réplication
- Hiérarchisation

NOSQL : Not Only SQL

- Cassandra
- Google's BigTable : HBase
- MongoDB (documents, JSON)
- CouchDB (documents, JSON)

Apache HADOOP

- HDFS (distributed high throughput FS)
- MapReduce
- HBase (scalable, distributed database)
- Hive (data warehouse infrastructure)
- Mahout (data mining library)
- Pig: (framework for parallel computation)
- ZooKeeper (distributed applications)

Amazon Elastic Map Reduce

- Logique de PaaS :
  - Stockage avec Amazon S3
  - Processing avec un cluster Amazon EC2
- Mise en oeuvre instantanée
- Simple
- Economique (0,3 $ / heure par node)

Database.com

- Database as a Service (DaaS)
- Multi-tenant
- Scalable à l'infini
- économique : ($10 / mois / 100000 records)

Conclusions

- Des solutions dans la continuité de l'existant.
- Des innovations permettant des ROI attrayants : Attention aux ruptures
- Outils opensource en voie de maturation

MERCI

Vincent Heuschling
vincent@heuschling.com
twitter : @vhe74

2.3 Alexandru Costan (INRIA)

Analyse des systèmes de stockage à grande échelle pour les applications de traitement intensif des données

Avec l'augmentation rapide des volumes de données dans de nombreux domaines d'application de la science de l'ingénierie et des services d'information, les défis posés par les traitements intensifs des données présentent une importance croissante. Avec l'émergence des infrastructures récentes (plates-formes de type *Cloud*, architectures massivement parallèles pétaflopiques) réaliser une gestion des données capable de passer à l'échelle dévient un enjeu crucial car les performances globales des applications dépendent des propriétés du service de gestion des données.

Nous définissons un ensemble de principes pour la conception de systèmes de stockage distribués, optimisés pour pouvoir passer à large échelle, tout en permettant autant de manipulations concurrentes des données que possible. Combinés, ces principes peuvent aider les développeurs de systèmes de stockage distribués à répondre aux exigences strictes de gestion de données à grande échelle.

Nous analysons ensuite plusieurs systèmes de stockage représentatifs afin d'évaluer la façon dont ils se conforment à ces principes de conception. Nous nous concentrons sur les systèmes de fichiers spécialisés qui ont été introduits pour cibler spécifiquement les besoins des applications de traitement intensif des données : `HDFS`, la couche de stockage standard utilisé par `Hadoop MapReduce`, `GPFS` proposé par `IBM`, ainsi que les systèmes de fichiers distribués massivement parallèles, comme `Lustre` ou `PVFS`, généralement utilisé pour les *clusters* de calcul à grande échelle. Avec l'émergence du calcul de type *Cloud*, des solutions de stockage spécialement conçues pour s'adapter à ce contexte ont été développés : nous présenterons texttttAmazon S3.

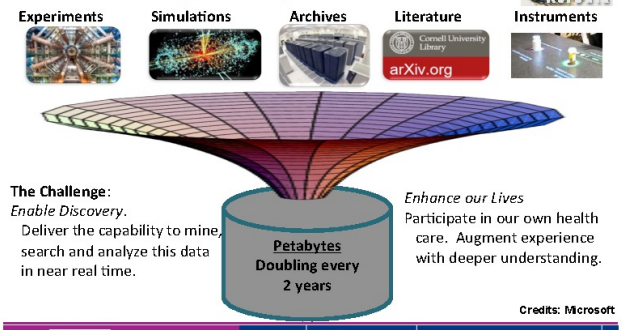
Nous détaillons en particulier les avantages potentiellement importants du versionnage pour améliorer les performances d'accès hautement concurrents aux données des applications. Dans ce contexte, nous proposons une interface d'accès basé sur la gestion de versions des données, matérialisée au sein de la plate-forme `BlobSeer` développée par l'équipe `KerData` de l'`INRIA` à Rennes. Cette approche qui permet d'exploiter d'une manière efficace le parallélisme inhérent des flux de données : nous en illustrons utilisation avec une application d'analyse conjointe de données génétiques et de neuro-imagerie.

A Survey of Large Scale Storage Systems for Data Intensive Applications

Alexandru Costan
KerData research team
INRIA Rennes - Bretagne Atlantique, France

Séminaire Aristote, Ecole Polytechnique, 9 June 2011

Context Today: an Explosion of Data



The Challenge:
Enable Discovery.
Deliver the capability to mine, search and analyze this data in near real time.

Enhance our Lives
Participate in our own health care. Augment experience with deeper understanding.

Credits: Microsoft

New Challenges for Large-scale Data Storage

Important issues:

- Scalable architecture (10^6 nodes)
- Massive unstructured data (Terabytes)
- Many data objects (10^9)
- Transparency
- High concurrency (10^3 concurrent clients)
- Fine-grain access (Megabytes)

Applications: distributed, with high-throughput requirements under concurrency

- Map-Reduce-based data-mining applications
- Governmental and commercial statistics
- Data-intensive HPC simulations
- Checkpointing for massively parallel computations
- On-Line Social Networks

Target platforms: from large clusters, grids and desktop grids to clouds and petascale machines

Big Data storage systems design principles

Data organization

- Scalability, transparency
- Ex: distributed files systems, object based storage devices (OSDs)

Asynchronous management

- Atomicity

Concurrency control

- Application-level parallelism
- Ex: locks, versioning

Data striping

- Configurable chunk distribution strategy
- Dynamically adjustable chunk sizes

Distributed metadata management

- Data availability

Specialized distributed storage systems

Data-intensive oriented file systems

- GFS
- HDFS

Parallel file systems

- GPFS
- Lustre

Cloud data storage services

- S3
- Azure

Specialized distributed storage systems

Data-intensive oriented file systems

- GFS
- HDFS

Parallel file systems

- GPFS
- Lustre

Cloud data storage services

- S3
- Azure

HDFS (Hadoop Distributed File System)

Part of Yahoo! Hadoop
- MapReduce implementation
- Open Source
- Java based

Distributed storage system
- Files are divided into large blocks (64 MB)
- Blocks are distributed across the cluster
- Blocks are replicated to help against hardware failure
- Data placement is exposed so that computation can be migrated to data

Notable differences from mainstream DFS work
- Single 'storage + compute' cluster vs. separate clusters
- Simple I/O centric API

HDFS Architecture: NameNode (1)

Master-Slave Architecture

HDFS Master "NameNode"
- Manages all file system metadata in memory
  - List of files
  - For each file name, a set of blocks
  - For each block, a set of DataNodes
  - File attributes (creation time, replication factor)
- Controls read/write access to files
- Manages block replication
- Transaction log: register file creation, deletion, etc.

HDFS Architecture: DataNodes (2)

HDFS Slaves "DataNodes"

A DataNode is a block server
- Stores data in the local file system (e.g. ext3)
- Stores meta-data of a block (e.g. CRC)
- Serves data and meta-data to Clients

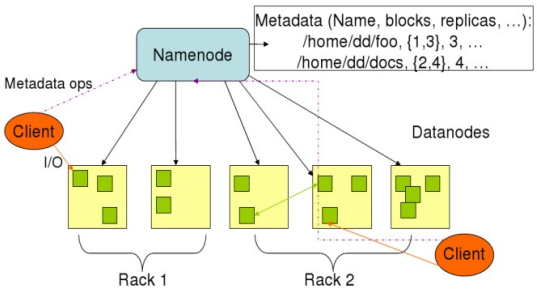
Block Report
- Periodically sends a report of all existing blocks to the NameNode

Pipelining of Data
- Forwards data to other specified DataNodes

Perform replication tasks upon instruction by NameNode

Rack-aware

HDFS Architecture (3)



Fault Tolerance in HDFS

DataNodes send heartbeats to the NameNode
- Once every 3 seconds

NameNode uses heartbeats to detect DataNode failures
- Chooses new DataNodes for new replicas
- Balances disk usage
- Balances communication traffic to DataNodes

Data corectness
- Use Checksums to validate data: CRC32

NameNode failures
- Single point of failures

Data-intensive oriented file systems

Huge files

Structured storage can be built on top

Fine grain concurrent reads

Pros
- No locking
- Data location aware

Cons
- Centralized metadata
- Expensive updates

Specialized distributed storage systems

Data-intensive oriented file systems
- GFS
- HDFS

Parallel file systems
- GPFS
- Lustre

Cloud data storage services
- S3
- Azure

GPFS (General Parallel File System)

Developed by IBM
- High-performance shared-disk clustered file system
- Used by many supercomputers in Top500

Distributed storage system
- Files are divided into *small blocks* (less than 1 MB)
- Blocks are distributed across the cluster
- Blocks are *RAID-replicated* or file system node replicated
- *Transparent* data location

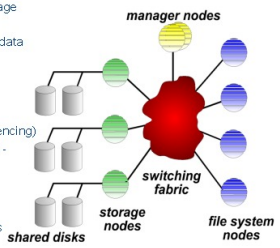
Notable differences
- *Distributed metadata*
- Efficient indexing of directory entries for very large directories.
- *POSIX semantics*
- Network partition aware

GPFS Architecture - Special Node Roles

File system nodes
- Run user programs, read/write data to/from storage nodes
- Cooperate with manager nodes to perform metadata operations

Storage nodes
- Implement block I/O interface
- Interact with manager nodes for recovery (e.g. fencing)
- Data and metadata striped across multiple disks - multiple storage nodes

Manager nodes
- File system configuration: recovery, adding disks
- Disk space allocation manager, quota manager
- File metadata manager - maintains file metadata integrity
- Global lock manager



Credits: IBM

Lustre

Massively parallel distributed file system (owned by Oracle)

Used by most supercomputers:
- The world's fastest computer - Tianhe-1A
- Jaguar (ORNL), LBNL, CEA

Features:
- OSD based
- Open source

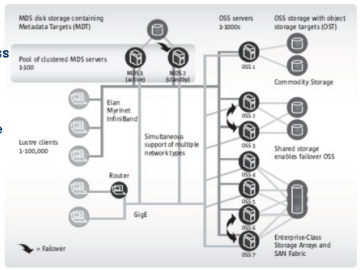
Lustre Architecture

Metadata Server (MDS)
- Active / Passive
- Filenames, directories, access permissions, file layout

Object Storage Servers (OSS)
- Store data on Object Storage Targets
- Distributed locking

Clients
- POSIX semantics

Fault tolerance: "failure as an exception"



Parallel file systems

Mounted as regular file systems

Data striping

Advanced caching

Pros
- Distributed data
- MPI optimizations

Cons
- Locking-based
- Too many small files

Specialized distributed storage systems

Data-intensive oriented file systems
- GFS
- HDFS

Parallel file systems
- GPFS
- Lustre

Cloud data storage services
- S3
- Azure

S3

Amazon Simple Storage Service:
- "storage for the Internet"
- (cheap) pay per use policy (for storage, requests, data transfers)

Design
- *Objects* (up to 5TB) stored in *buckets*, identified using *keys*
- Buckets stored in one of several Regions
- Clients authorization using ACLs
- Access through Web interfaces: REST, SOAP, BitTorrent

Notable uses
- FUSE – allows EC2-hosted Xen images mount an S3 bucket as a file system
- Apache Hadoop
- Tumblr

Azure

Proposed by Microsoft within Windows Azure PaaS cloud

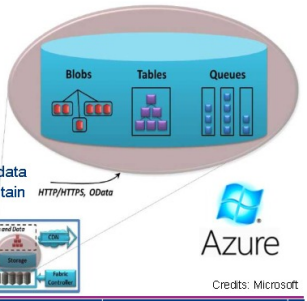
Data manipulation based on HTTP

All data replicated 3 times

Blobs
- Up to 1 TB of unstructured data
- Grouped in containers

Tables
- Fine grained access to structured data
- Group of entities / records that contain properties

Queues
- Asynchronous comm. between cloud instances



Credits: Microsoft

Cloud data storage services

Virtualize storage resources

Pay for duration, size and traffic

Flat naming scheme

Simple access model

Pros
- High data availability
- Versioning

Cons
- Limited object size
- Low throughput

Limitations of existing approaches

Issue	Parallel FS	Data-intensive FS	Cloud store
Too many small files	✖	✔	✖
Centralized metadata	✔	✖	✔
No versioning support	✖	✖	✔
No fine grain writes	✔	✖	✖

✔ = addressed issue

[Nicolae et al., 2010]

Concurrency-optimized BLOB management: The BlobSeer Approach

BlobSeer: software platform for scalable, distributed BLOB management
- Huge data (TB) - BLOBs: Binary Large OBjects
- Highly concurrent, fine-grain access (MB): Read/Write/Append
- Developed by the KerData team at INRIA, Rennes

Overview of key design choices
- Decentralized data storage
- Decentralized metadata management
- Versioning-based concurrency control, multiversioning exposed to the user
- Lock-free concurrent writes (enabled by versioning)

A back-end for higher-level, sophisticated data management systems
- Short term: highly scalable distributed file systems
- Middle term: storage for cloud services
- Long term: extremely large distributed databases

http://blobseer.gforge.inria.fr/

BlobSeer: Key Design Choices

Distributed data

- Each BLOB is fragmented into "chunks" (pages)
- Huge data amounts to be distributed all over the storage nodes
- Reduced contention for simultaneous accesses to disjoint parts of the BLOB

Distributed Metadata

- Goal: locate chunks that make up a given BLOB
- Fine-grained and distributed
- Efficiently managed through a segment tree over a DHT

Versioning-based concurrency control

- Update/append: generate new chunks rather than overwrite
- Metadata is extended to incorporate the update
- Both the old and the new version of the BLOB are accessible
- Lock-free approach

http://blobseer.gforge.inria.fr/

BlobSeer: Architecture

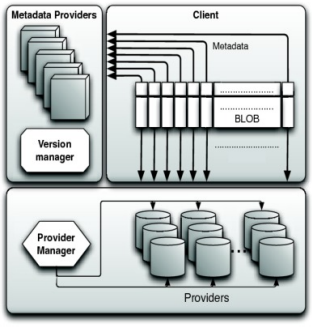
Clients
- Perform fine grain BLOB accesses

Providers
- Store the chunks of the BLOB

Provider manager
- Monitors the providers
- Favors data load balancing

Metadata providers
- Store information about chunk location

Version manager
- Ensures concurrency control



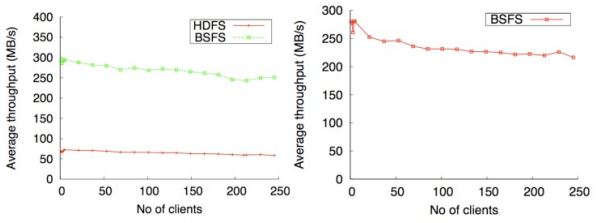
Integrating BlobSeer in the Hadoop Map-Reduce Framework

MapReduce: a natural application class for BlobSeer:
- Case study: Yahoo!'s Hadoop MapReduce framework
- Approach: use BlobSeer instead of Yahoo!'s Hadoop file system (HDFS)
- Motivation: HDFS has limited support for concurrent access to shared data

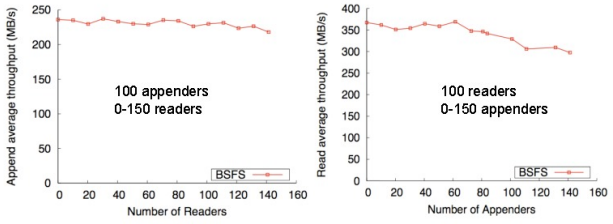
Implementing the HDFS API for BlobSeer
- Implements basic file system operations: create, read, write…
- Introduces support for concurrent append operations

BlobSeer File System (BSFS)
- File system namespace - keeps file metadata, maps files to BLOB's
- Client-side buffering: data prefetching, write aggregation
- Exposes data layout to Hadoop, just like HDFS

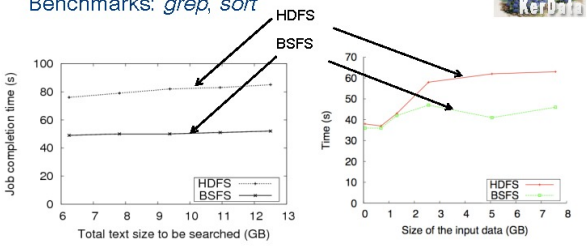
BSFS vs. HDFS Concurrent Reads, Concurrent Appends



BlobSeer Scales Up: Readers and Writers Do Not Interfere (Almost!)



BSFS Does Better Than HDFS! Benchmarks: *grep*, *sort*



Relevant publications
- JPDC (2010), Special Issue on Data-Intensive Computing
- IEEE IPDPS 2010
- MapReduce 2010 (held in conjunction with HPDC 2010)

The AzureBrain Project: BlobSeer on Microsoft Azure Clouds

Application
- Large-scale Joint Genetic and Neuroimaging Data Analysis

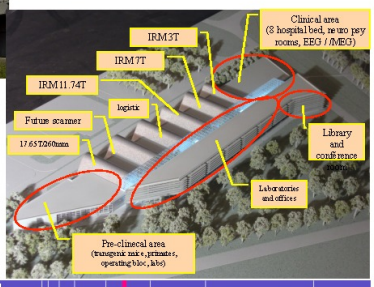
Goal
- Assess and understand the variability between individuals

Approach
- Optimized data processing on Microsoft's Azure clouds based on the BlobSeer concurrency-optimized platform

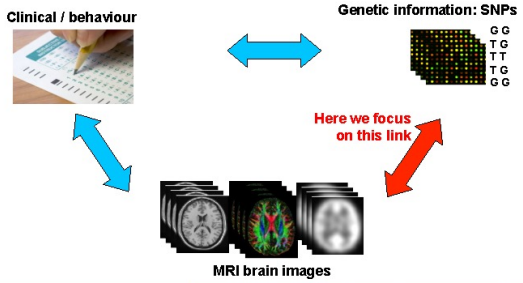
INRIA teams involved
- KerData (Rennes)
- PARIETAL (Saclay)

Framework
- Joint MSR-INRIA Research Center
- MS involvement: Azure teams, EMIC

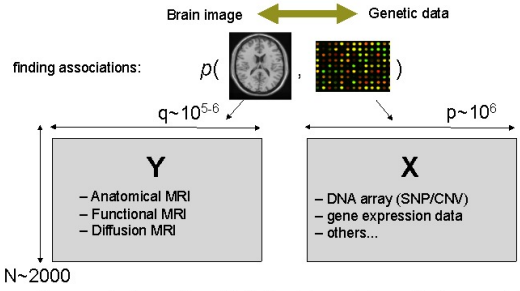
NeuroSpin
Neuroimaging center at Saclay



The Imaging Genetics Challenge: Comparing Heterogeneous Information

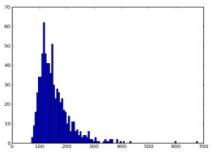


Imaging Genetics Methodological Issues

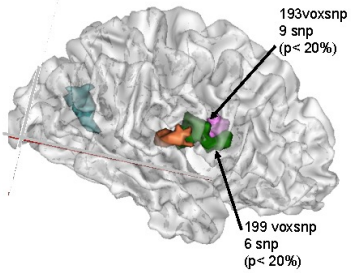


Approach: Searching Statistical Associations Between Pairs

Illustration



4D cluster size d distribution under H0

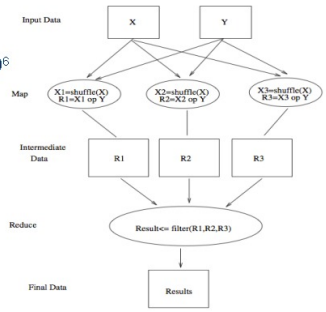


A-Brain:
The goal is to reproduce this kind of study with 10^5 larger data

The Computational Problem

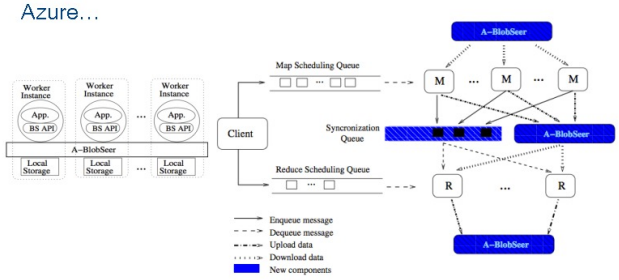
- Neuroimaging data (voxels in each contrast map): 10^5 to 10^6
- Genetic data: 10^6 variables
- Permutation tests: 10^3

Around 10^{15} tests

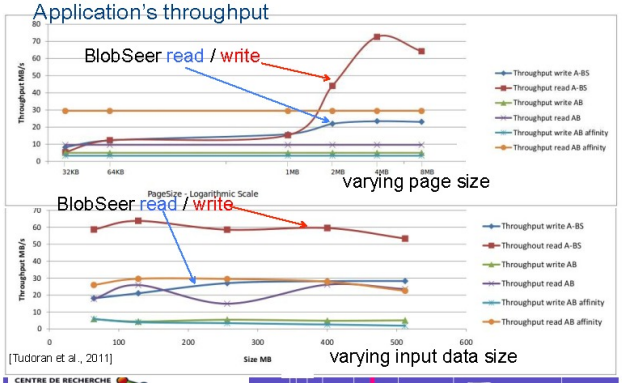


BlobSeer as a storage backend in Azure…

… used within the MapReduce based application



Application's throughput



Summary

Difficult to maximize all the objectives: achieve a very high data throughput for highly concurrent, fine-grain data accesses

Concurrency control based on locking mechanisms often creates bottlenecks

Object based storage approaches ensure scalability

Consistency model: CAP

Data-intensive specific solutions exploit application level parallelism but force users to adhere to a specific programming paradigm

Thank you!

For more information…

- *BlobSeer: http://blobseer.gforge.inria.fr*
- *KerData: http://irisa.fr/kerdata*

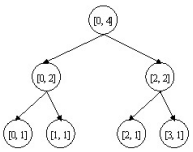
Metadata Zoom (1)

Organized as a segment tree

Each node covers a range of the blob identified by (offset, size)

The first/second half of the range is covered by the left/right child

Each leaf corresponds to a chunk and holds information about its location



Metadata Zoom (2)

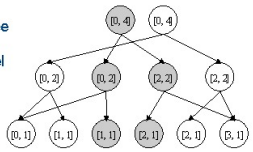
Each node holds versioning Information

Write/Append
- Add leaves and build subtree up to the root
- The tree may grow one level

Read: descend from the root towards the leaves

Tree nodes are distributed among metadata providers

Highly scalable access concurrency: R/R, R/W, W/W



Metadata Zoom (2)

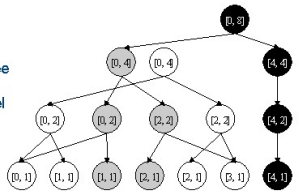
Each node holds versioning Information

Write/Append
- Add leaves and build subtree up to the root
- The tree may grow one level

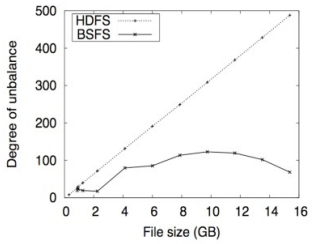
Read: descend from the root towards the leaves

Tree nodes are distributed among metadata providers

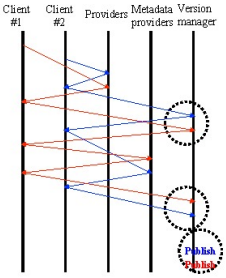
Highly scalable access concurrency: R/R, R/W, W/W



BSFS vs. HDFS
Load balancing the storage nodes



How Versioning Enables Efficient, Heavy Access Concurrency



Chunks are written concurrently by the clients

Then, versions are assigned in the order the clients finish writing

Then, metadata is written concurrently by the clients

Versions are published in the order they were assigned

Leveraging BlobSeer on Clouds: MapReduce

MapReduce: a simple programming model for data-intensive computing

Typical problem solved by MapReduce
- Read a lot of data
- Map: extract something you care about from each record
- Shuffle and Sort
- Reduce: aggregate, summarize, filter, or transform
- Write the results

Approach: hide messy details in a runtime library
- Automatic parallelization
- Load balancing
- Network and disk transfer optimization
- Transparent handling of machine failures

Implementations: Google MapReduce, Hadoop (Yahoo!)

2.4 Joerg Bienert (ParStream)

An innovation solution to manage heterogeneous big data

Analyzing Big Data is getting more and more important for companies in all industries, *e.g.* web analytics, fraud detection, smart metering, telco, *etc.* Current established database products are not able and not designed to perform large scale data analytics. New approaches like Map/Reduce lack important features like short response times. `ParStream` is a novel inovative database technology focusing on processing large data sets (billions of records) in millisecondes and with low latency. `Parstream` is a columnar in memory database running on multiprocessor architectures and, as first product, on `GPU` based `HPC`-Servers.

ParStream

Big Data – Real Time – Low Latency

Ecole Polytechnique

June 9th 2011

Agenda

ParStream

Big Data – The Challenge

Current Approaches

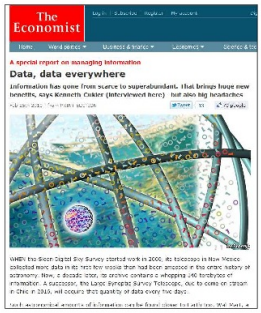
ParStream – A HPC Database for Big Data

Use Cases

Big Data

ParStream

The amount of digital information increases tenfold every five years



- "We are at a different period because of so much information", says James Cortada of IBM
- "..what is scarce is the **ability to extract wisdom** from them.", Hal Varian, Google's chief economist
- "Every day I wake up and ask, how can I flow data better, manage data better, analyse data better?", says Rollin Ford, the CIO of Wal-Mart.
- "The **data-centered economy** is just nascent", admits Mr. Mundie of Microsoft.

Big Data

ParStream

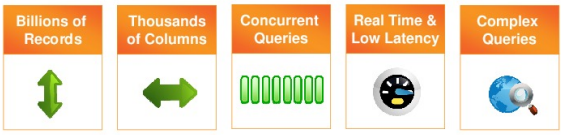
„Analyzing big data—will become a key basis of competition, underpinning new waves of productivity growth, innovation, and consumer surplus"

- Making big data more accessible in a timely manner
- Using data and experimentation to expose variability and improve performance
- Segmenting populations to customize actions
- Replacing and supporting human decision-making with automated algorithms
- Innovating new business models, products, and services.

Challenges in Big Data Analytics

ParStream

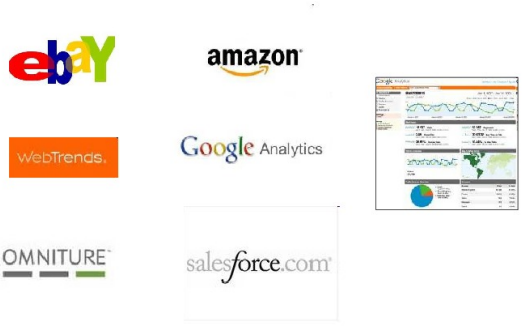
There are challenges in Big Data Analysis in different dimensions



Big Data

ParStream

eCommerce & Web Analytics



Big Data

ParStream

Social Media Analytics

facebook

twitter

foursquare

LinkedIn

7

Big Data in Telecommunications

ParStream

Billing, Pricing, Fraud detection

8

Big Data in Utilities

ParStream

Smart meetering & Smart grid

9

Big Data in Finance

ParStream

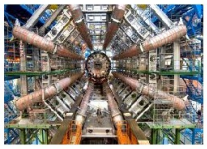
Share Prices History, Fraud Detection, Algo-Trading

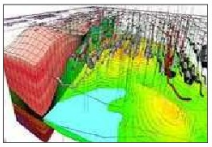
10

Big Data in other Industries

ParStream

Retail, Telematics, Science, Mining





11

Established Databases

ParStream

Current databases are not engineered for mass-data

"Existing Database Architectures are 20-30 years old and are not able to cope with current data sizes."

Several Statements on Conference for Very Large Databases VLDB 2010, Singapore

12

Hadoop Ecosystem

ParStream

Some New approaches not suitable for Realtime Processing

"MapReduce isn't suited to calculations that need to occur in near real-time. You can't do anything that takes a relatively short amount of time, so we got rid of it."

Lipkovitz, Senior Director of Engineering, Google

13

Big Data Threat

ParStream

Big Data Analysis requires new "engines"

14

Big Data products

ParStream

New Marketplayers – and their downsides

No Partitioning — VERTICA

No Index, Postgres based — Greenplum

Map Reduce approach — aster data

Hardware bound — NETEZZA

No Index — EXASOL

No Index — PARACCEL

15

Big Data

ParStream

New Market Players – and their setbacks

No Partitioning — hp VERTICA

No Index, Postgres based — EMC² Greenplum

Map Reduce approach — TERADATA aster data

Hardware bound — IBM NETEZZA

No Index — EXASOL

No Index — PARACCEL

16

ParStream

ParStream

ParStream

The first HPC Database focusing on the Real-Time Analysis of Big Data

17

ParStream – Building Blocks

ParStream

ParStream combines state of the art database technologies with unique technologies

Column Store	In-Memory Technology	High Performance Index	Custom Query Operators
Fast data access for analytical processing	Data and indices can stay in memory due to efficient compression	Unique index structure allows highly parallel execution of queries	SQL, JDBC and powerful C++ API enable fast query processing

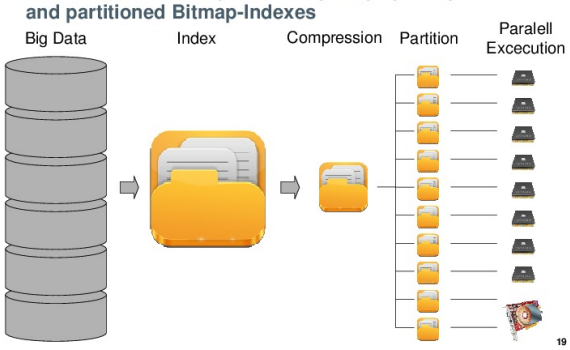
Patent Pending – High USP

18

High Performance Indexing

ParStream

Fast and flexible analytics through highly compressed and partitioned Bitmap-Indexes

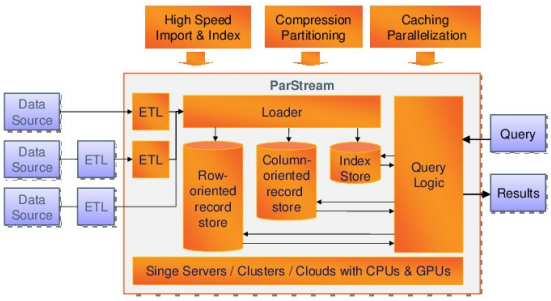


19

High Level Technical Architecture

ParStream

ParStream is a complete database on a highly optimized architecture



20

ParStream - Product Features

ParStream

ParStream is optimized for all challenges in Big Data analysis

Queries include new data immediately after continuous import

Configurable partitioning allows horizontal and vertical clustering (fault tolerant)

ParStream scales linearly up to PetaByte

SW Package for Linux on Standard HW (Windows upc.)

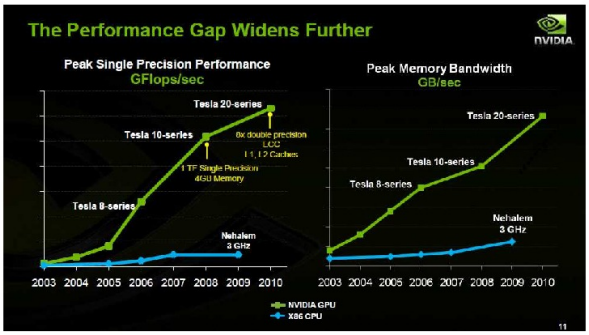
GPU Server Appliance

21

GPU Processing – motivation

ParStream

Graphical Processing Units (GPUs) outperform standard processors



11

Cloud Ready

ParStream

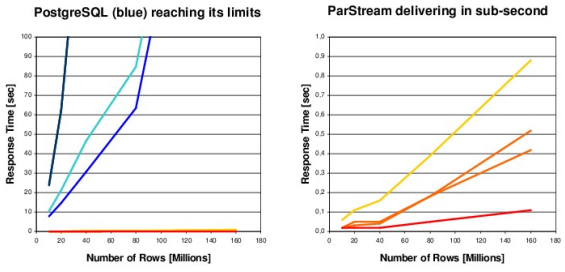
ParStream delivers outstanding performancen all infrastructure Setups

23

Performance ParStream vs PostgreSQL

ParStream

ParStream delivers in sub-seconds even on > 100 million rows
ParStream clearly dominates PostgreSQL and others



24

Explorative analytics on billions of records

ParStream

ParStream analyzes billions of records within milliseconds

Case: Web Analytics

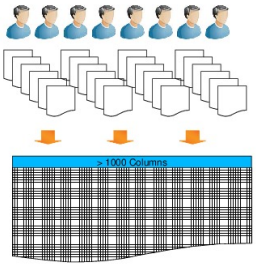
- Unique Visitor Calculation (select count distinct)
- Analysis of 8 billion records in 15 ms on 4-server cluster

25

Flexible analytics on thousands of columns

ParStream

ParStream allows quering data by selecting any combination of different columns



Case: Market Research

- Pattern recognition
- Flexible multi-column filtering & grouping
- 20 million records with 1000 data columns
- 5000 queries in 5 seconds

26

Concurrent queries and high throughput

ParStream

ParStream can execute many queries efficiently at the same high throughput

Case: Online Travel Search

- High throughput even with complex filter & sorting criteria
- 100 queries per second per node on 1 billion travel offers
- Throughput scales linearly with number of nodes

27

Use case: travel search engine

ParStream

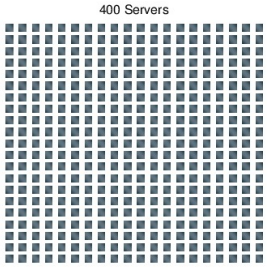
Comparison for a search appliance delivering 100 travel offers out of 1 billion based on 25 independent, optional filter criteria

28

Use case: travel search engine

ParStream

Server Requirements



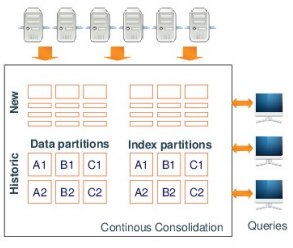
Sizing for a search appliance executing 1000 queries per second to deliver 100 travel offers out of 1 billion based on 25 optional filter criteria

29

Instant analytics on up-to-the-second data

ParStream

Simultaneous data import and querying allows comparison of current & historical data



Case: Web Analytics

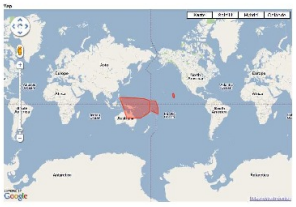
- Continuous import of newly created data every second
- Simultaneous import and querying
- Single node performance typically 1TB per hour
- Continuous trend spotting for alert-function

30

Advanced analytics

ParStream

Standard SQL features and custom query nodes provide maximum flexibility and speed



Case: Climate Research

- Identification of hurricane risk areas
- Geo-clustering with customized query node
- Implementation of custom algorithms via C++ API
- Analysis of 3 billion records in 100 ms

31

Testimonials

ParStream

Leading experts, customers and hardware providers have realized the benefits of ParStream

Experts are impressed…

- "An extremely innovative idea"
- "Convinced me completely"
- "2 years ahead of competition"

Prof. Dr. Markl,
leading database expert at TU Berlin,
Formerly at IBM for DB2 optimization

Customers are satisfied…

- Outperforms all databases by a factor of at least 35
- Reduced our response time by factors up to 12.000
- Scales linearly, runs stable in production

Leading web-analytics company

Awards…

- "One to watch award 2010" from Nvidia, Cooley & Adobe
- "First database running on high-performance GPUs"

32

Summary

ParStream

- Analyzing Big Data is getting more and more important for every industry
- Existing Database Architectures are 25-30 years old and not designed for Big Data
- New approaches often do not fulfill promises
- ParStream is a new powerful database, specifically designed for Big Data Analysis

33

Contact

ParStream

Questions ?

joerg.bienert@parstream.com
Phone +49 221 97 76 14 80

Trusted Cloud

34

2.5 Ronan Keryell (HPC-Project)

Environnement de programmation pour traitements massifs sur architectures modernes

L'ubiquité de l'informatique déclenche une avalanche de données à traiter de manière rapide et économique. Malheureusement, pour des raisons physiques, la loi de Moore ne fournit plus de processeurs plus rapides (dissipation thermique, vitesse de la lumière, ...) mais fournit néanmoins toujours plus de transistors. Le seul moyen d'utiliser ces transistors est d'utiliser du parallélisme massif, mais cela remet en question les architectures et les modes de programmation. Les architectures actuelles (`GPU`, `MP-SoC`, `Tilera`, `FPGA`, ...) seront présentées avec leurs avantages et leurs contraintes ainsi que leurs pendants logiciels permettant de les utiliser au mieux. L'environnement de compilation source à source `Par4All` de `HPC Project` est un moyen de s'abstraire de certains détails de programmation.

Programming environments for big data processing on modern parallel architectures

Ronan KERYELL

HPC Project

2011/06/09

—

Le déluge de données, comment en tirer parti ?
Séminaire Aristote

HyperParallel Technologies (1992–1998)

- Parallel computer
- Proprietary 3D-torus network
- DEC Alpha 21064 + FPGA
- HyperC (follow-up of PompC @ LI/ENS Ulm)
  - PGAS (Partitioned Global Address Space) language
  - An ancestor of UPC...
- Already on the Saclay Plateau ! ☺

Quite simple business model

- Customers need just to rewrite all their code in HyperC ☺
- Difficult entry cost... ☹

- Niche market... ☹
- American subsidiary with dataparallel datamining application acquired by Yahoo! in 1998
- Closed technology ⇝ lost for customers and... founders ☹

HyperParallel Technologies (1992–1998)

- Parallel computer
- Proprietary 3D-torus network
- DEC Alpha 21064 + FPGA
- HyperC (follow-up of PompC @ LI/ENS Ulm)
  - PGAS (Partitioned Global Address Space) language
  - An ancestor of UPC...
- Already on the Saclay Plateau ! ☺

Quite simple business model

- Customers need just to rewrite all their code in HyperC ☺
- Difficult entry cost... ☹

- Niche market... ☹
- American subsidiary with dataparallel datamining application acquired by Yahoo! in 1998
- Closed technology ⇝ lost for customers and... founders ☹

Programming environments for big data processing on modern parallel architectures
HPC Project Ronan KERYELL 2 / 101

Present motivations: reinterpreting Moore's law *(I)*

The good news ☺

- Number of transistors still increasing
- Memory storage increasing (DRAM, FLASH...)
- Hard disk storage increasing
- Processors (with captors) everywhere
- Network is increasing

- The bad news ☹
  - Transistors are so small they leak... Static consumption
  - Superscalar and cache are less efficient compared to transistor budget
  - Storing and moving information is expensive, computing is cheap: change in algorithms...
  - Light's speed as not improved for a while... Hard to reduce latency
    - Chips are too big to be globally synchronous at multi GHz ☹
  - pJ and physics become very fashionable

Present motivations: reinterpreting Moore's law *(II)*

- Power efficiency in $\mathcal{O}(\frac{1}{7})$
  - Transistors cannot be used at full speed without melting ☹ 🖷
  - Heat
- I/O and pin counts
  - Huge time and energy cost to move information outside the chip ☹
- Rotating hard disk with 1D density d increase
  - Storage in $\mathcal{O}(d^2)$
  - But track speed only $\mathcal{O}(d)$

⇝ Reading all the disk in $\mathcal{O}(\frac{1}{d})$ ☹

Heterogeneous parallelism

Parallelism is the only way to go...

- Research is just crossing reality!
- Scaring... ☹
- Exciting! ☺

No one size fit all...

Future will be heterogeneous

The “Software Crisis”

Edsger DIJKSTRA, 1972 Turing Award Lecture, « The Humble Programmer »

“To put it quite bluntly: as long as there were no machines, programming was no problem at all; when we had a few weak computers, programming became a mild problem, and now we have gigantic computers, programming has become an equally gigantic problem.”

`http://en.wikipedia.org/wiki/Software_crisis`

⚠ But... it was before parallelism democratization! ☺

Time to be back in parallelism!

- Good time for more start-ups! ☺
- ⇝ 2006: thinking to yet another start-up...
- People that met ≈ 1990 at the French Parallel Computing military lab SEH/ETCA
- Later became researchers in Computer Science, CINES director and ex-CEA/DAM, venture capital and more: ex-CEO of Thales Computer, HP marketing...
- HPC Project launched in December 2007
- ≈ 30 colleagues in France (Montpellier, Meudon), Canada (Montréal with Parallel Geometry) & USA (Mountain View)

HPC Project hardware: WildNode from Wild Systems

Through its Wild Systems subsidiary company

- WildNode hardware desktop accelerator
  - Low noise for in-office operation
  - *x*86 manycore
  - nVidia Tesla GPU Computing
  - Linux & Windows

`http://www.wild-systems.com`

HPC Project software and services

- Parallelize and optimize customer applications, co-branded as a bundle product in a WildNode (e.g. Presagis Stage battle-field simulator, WildCruncher for Scilab//...)
- Acceleration software for the WildNode
  - GPU-accelerated libraries for C/Fortran/Scilab/Matlab/Octave/R
  - Transparent execution on the WildNode
- Remote display software for Windows on the WildNode

HPC consulting

- Optimization and parallelization of applications
- *High Performance?...* not only TOP500-class systems: power-efficiency, embedded systems, green computing...
- ⇝ Embedded system and application design
- Training in parallel programming (OpenMP, MPI, TBB, CUDA, OpenCL...)

Efficient big data architectures (I)

- Massive parallelism
- Use right processing elements for the right tasks (efficiency)
- Avoid moving data (expensive, slow)
- Avoid storing data: on-the-fly processing of data to be correlated
- Use memory hierarchies
- Distributed computing with smart routers

Outline

Outline

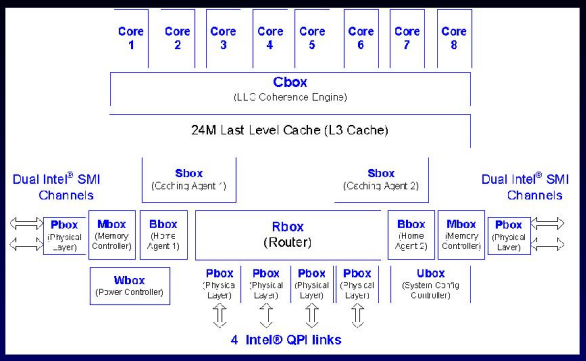
Intel Nehalem (I)

- x86-64 new microarchitecture
- With on-chip memory controllers like AMD Opteron: ↗ bandwidth, ↘ latency
- SSE4.2 instruction set: strings (XML parser...), comparisons (data-mining), CRC & cryptography (protocols)
- Power consumption fine tuning with many sensors (power, temperature). Possible to speed up when less cores are used (turbo)
- Xeon X7560 (Beckton microarchitecture), 2010
  - 8 cores @ 2.27 GHz (+ turbo 2.666 GHz) + SMT 2 threads (HyperThreading) with 256 KB L2 cache/core, 32D+32I KB cache/core
  - 24 MB L3 cache
  - 4 Quick Path Interconnects (QPI) @ 6.4 GT/s to play Lego with processors & accelerators (≈HyperTransport d'AMD)

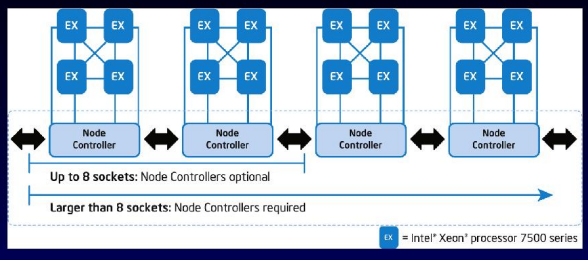
Intel Nehalem (II)

- 4 DDR3-1333 MHz memory controllers
- 130 W
- $3692 March 30, 2011
- 2.3 Gtr 45 nm 684 mm^2

Intel Nehalem (III)



Intel Nehalem (IV)



AMD Opteron 6180 (2011)

- x86-compatible instruction set
- 64-bit mode that double also register numbers
- Out-of-order superscalar execution with 9 instructions/cycle
  - 3 integer instructions
  - 3 address generators
  - 3 floating computation operators (+, *, memory)
- Opteron 6180 SE, 02/2011, $1514
  - 12 cores @ 2.5 GHz, 512 KB/core L2 cache
  - 2 × 6 MB L3 cache
  - 21 GB/s DDR3 memory
  - 4 x 6.4 GT/s HyperTransport
  - 1.8 Gtr 45 nm 692 mm^2
  - 140 W

Hardware architectures ▶ Classical multicores

IBM Power 7 (2010)

- 4 chips per quad-chip module
- 8 cores per chip @ 4.0 GHz (4.25 GHz in TurboCore mode with 4 cores)
- 1.2 Gtr 45 nm SOI process, 567 mm^2
- 4 SMT threads per core
- 32I+32D kB L1 cache/core
- 256 kB L2 cache/core
- 32 MB L3 cache in eDRAM
- 100 GB/s DDR3 memory
- 12 execution units per core:
  - 2 fixed-point units
  - 2 load/store units
  - 4 double-precision floating-point units
  - 1 vector unit supporting VSX
  - 1 decimal floating-point unit
  - 1 branch unit
  - 1 condition register unit

Hardware architectures ▶ GPU

Outline

Hardware architectures ▶ GPU

Off-the-shelf AMD/ATI Radeon HD 6970 GPU

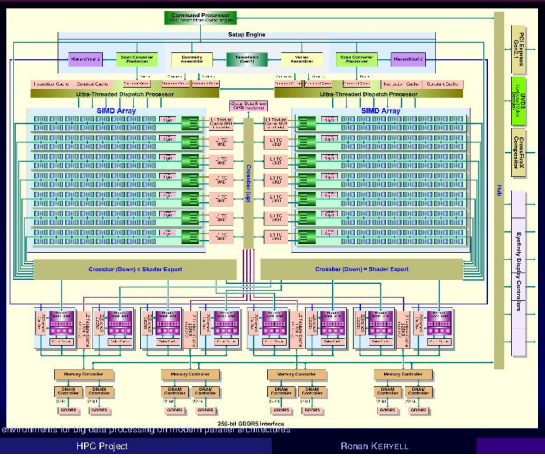
- 2.64 billion 40nm transistors
- 1536 stream processors @ 880 MHz, 2.7 TFLOPS SP, 675 GFLOPS DP
- + External 1 GB GDDR5 memory 5.5 Gt/s, 176 GB/s, 384b GDDR5
- 250 W on board (20 idle), PCI Express 2.1 x16 bus interface
- OpenGL, OpenCL
- ∃ Radeon HD 6990 double chip card

More integration:

- Llano APU (FUSION Accelerated Processing Unit) : x86 multicore + GPU 32nm, OpenCL

Hardware architectures ▶ GPU

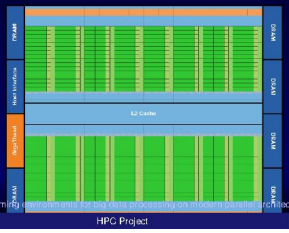
Radeon HD 6870 — big picture

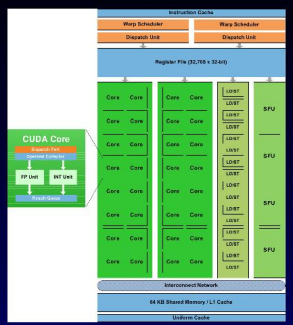


Hardware architectures ▶ GPU

Off-the-shelf nVidia Tesla Fermi M2090 & GTX580

- GF110: 3 billion 40nm tr.
- 512 thread processors @ 1300 MHz, 1,3 TFLOPS SP, 666 GFLOPS DP
- + External 6 GB GDDR5 ECC memory 3,7 Gt/s, 177 GB/s. Less if using ECC

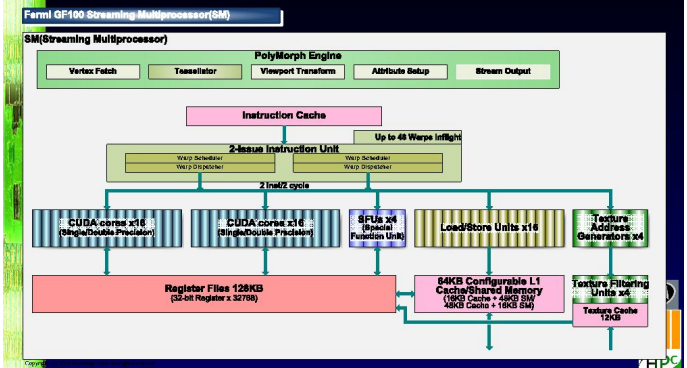




- 247 W on board PCI Express 2.1 x16 bus interface
- OpenGL, OpenCL, CUDA

Hardware architectures ▶ GPU

GF100 Stream Multiprocessor



Hardware architectures ▶ MP-SoC underworlds

Outline

Hardware architectures ▶ MP-SoC underworlds

ARM yourself (I)

- Do some computations where the captors are...
- Smartphone and other sensor networks
- Trade-off between communication energy and inside/remote computations
- Texas Instrument OMAP4470 announced on 2011/06/02
  - 2 ARM Cortex-A9 MPCores @ 1.8GHz with Neon vector instructions
  - 2 ARM Cortex-M3 cores (low-power and real-time responsiveness, multimedia, avoiding to wake up the Cortex-A9...)
  - **SGX544 graphics core with OpenCL 1.1 support**, with 4 USSE2 core @ 384 MHz producing each 4 FMAD/cycle: 12.3 GFLOPS
  - 2D graphics accelerator
  - 3 HD displays and up to QXGA (2048x1536) resolution + stereoscopic 3D
  - Dual-channel, 466 MHz LPDDR2 memory

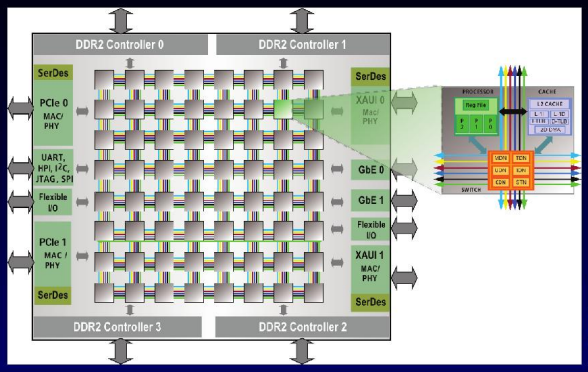
Hardware architectures ▶ MP-SoC underworlds

ARM yourself (II)

- ⇝ Current course to have non-x86 servers based on ARM...
- ∃ Experiments on low power clusters
- Think to evaluate power consumption on your application

Hardware architectures ▶ MP-SoC underworlds

Tilera TilePro64 (I)



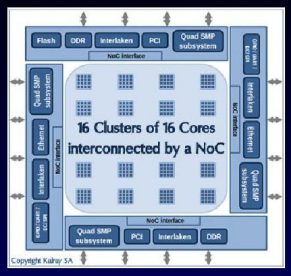
Hardware architectures ▶ MP-SoC underworlds

Tilera TilePro64 (II)

- Processing power with good network interfaces
  - Interactive data analysis
  - Video/audio codec
  - DPI, IDS, IPsec...
- 8x8 processors with SMP, partitionable
- 32-bit VLIW with SIMD mode
- 700-866 MHz: 443 GOPS 8 bits, 23 W
- 64 DDR2 controllers 25.6 GB/s
- 2× 10 GbE XAUI + IP session hash distribution
- SMP Linux or bare bone per tile
- C/C++ gcc compiler
- TMC library for hardcore support (2D network...)
- Eclipse support with graphical simulator
- OpenMP

Hardware architectures ▶ MP-SoC underworlds

MPPA de Kalray (the French touch!) (I)



- 256 VLIW processors per chip 16 clusters of 16 processors)
- Shared memory and NoC with DMA
- 28 nm CMOS technology, ≈ 5 W @ 400 MHz
- FPU 32/64 bits IEEE 754: 205 GFLOPS SP, 1 TOPS 16 bits
- 2× 64-bit DDR3 memory controllers for high bandwidth main memory transfers
- 2× 40 Gb/s or 8× 10 Gb/s) Ethernet controller

MPPA de Kalray (the French touch!) (II)

- 2× 8-lane PCI Express Gen 3
- 4× 4–8-lane Interlaken interfaces for multi-MPPA chip system integration (8 MPPA/PCIe board) or connection to external FPGAs, I/O...
- Linux or bare metal with AccessCore library
- Multi-core compiler (gcc 4.5), simulator, debugger (gdb), profiler
- Eclipse IDE
- Programming from high-level C-based language
- AccessCore library

FPGA (I)

- Field-programmable gate array with bitstream configuration in memory
- Xilinx Virtex 7
  - 2M logic cells (6-LUT), 28 nm
  - 85Mb block RAM
  - 5280 DSP slices (6.7 TFMA/s)
  - 96 transceiver @ 28Gb/s: 2.8 Tb/s
  - PCIe gen3 ×8
  - 1200 pins
- Radar, communications, HPC, datamining, bioinformatics...

FPGA (II)

- VHDL-to-bitstream compiler... but **hard work**
- Dynamic partial reconfiguration
- ∃ C-to-VHDL compilers
  - Riverside Optimizing Compiler for Configurable Computing (ROCCC): open source
  - Impulse C
  - Catapult C
  - Cynthesizer
  - ...

Convey HC-1^{ex}

- Intel Xeon quad-core @2.13 GHz with 128 GB
- 4 Xilinx Virtex 6 LX760 FPGA with 128 GB DDR2
- 1520 W in 3U rack
- Linux
- Various instruction sets (“personality”)
- Fortran/C/C++ vectorizing compiler replacing complex instructions by FPGA vectorized implementation
- Possible to design its own instruction set (ROCCC...)
- 401× speed-up on Smith-Waterman algorithm compared to *x*86

Anton computer from D.E. Shaw

1. Create a hedge fund
2. Earn a lot of money
3. Spend it by creating a 500+ people start-up in bioinformatics
4. Build from scratch (ASIC) a computer for solving special issues

PacketShader

- Use Linux PC + GPU as a router with processing power
  - 2 4-core Nehalem @2.66 GHz + 4 10 GigE NIC + 2 GPU GTX480
- 40 Gb/s routing even with 64-byte packets
- 8–20 Gb/s IPsec tunneling
- Linux IP stack to slow ⇝ own raw device driver for Intel 82598/82599 NIC
  - Extract Ethernet flow into **user** mode
  - Huge recycled packet buffers
  - Batch processing to amortize overhead
  - NUMA-aware processing: packets processed by NIC-local CPU in its RAM, aligned in cache
- Current bottleneck: IOH chipset

`http://shader.kaist.edu/packetshader`

OpenFlow

- Need for advanced routing with computing capabilities
- PC or Tilera-like with few 10+ Gb/s Ethernet or Infiniband: bandwidth maybe OK but not enough links
- ⇝ Open standard to interact with existing routers & switch: OpenFlow `http://www.openflow.org`
- Possible to extract *minimal* flows to feed PC/GPU/MP-SoC/FPGA accelerators and reinject results in the router
- OpenFlow implemented by many router companies
- Possible with non-OpenFlow but less flexible and non portable

Outline

Outline

Parallel application dwarfs

- `http://view.eecs.berkeley.edu/` : « The Landscape of Parallel Computing Research: A View From Berkeley »
- Try to capture typical examples to analyze and design new architectures & applications

	Dwarf	Performance Limit: Memory Bandwidth, Memory Latency, or Computation?
1	Dense Matrix	Computationally limited
2	Sparse Matrix	Currently 50% computation, 50% memory BW
3	Spectral (FFT)	Memory latency limited
4	N-Body	Computationally limited
5	Structured Grid	Currently more memory bandwidth limited
6	Unstructured Grid	Memory latency limited
7	MapReduce	Problem dependent
8	Combinational Logic	CRC problems BW; crypto problems computationally limited
9	Graph traversal	Memory latency limited
10	Dynamic Programming	Memory latency limited
11	Backtrack and Branch+Bound	?
12	Construct Graphical Models	?
13	Finite State Machine	Nothing helps!

Extracting parallelism in applications...

- The implicit attitude
  - Hardware: massively superscalars processors
  - Software: auto-parallelizing compilers
- The (±) explicit attitude
  - Languages (± extensions): OpenMP, UPC, HPF, Co-array Fortran (F--), Fortran 2008, X10, Chapel, Fortress, Matlab, SciLab, Octave, Mapple, LabView, nVidia CUDA, AMD/ATI Stream (Brook+, Cal), OpenCL, HMPP, *insert your own preferred language here* ...
  - Framework: MapReduce, Hadoop...
  - Libraries: application-oriented (mathematics, coupling...), parallelism (MPI, concurrency *pthreads*, SPE/MFC on Cell...), Multicore Association MCAPI, objects (parallel STL, TBB, Ct...)

... but multidimensional heterogeneity!

Welcome into Parallel Hard-Core Real Life 2.0!

- Heterogeneous execution models
  - Multicore SMP ± coupled by caches
  - SIMD instructions in processors (Neon, VMX, SSE4.2, 3DNow!, LRBni...)
  - Hardware accelerators (MIMD, MISD, SIMD, SIMT, FPGA...)
- New heterogeneous memory hierarchies
  - Classic caches/physical memory/disks
  - Flash SSD is a new-comer to play with
  - NUMA (*Non Uniform Memory Access*) : sockets-attached memory banks, remote nodes...
  - Peripherals attached to sockets : NUPA (*Non Uniform Peripheral Access*). GPU on PCIe ×16 in this case...
  - If non-shared memory: remote memory, remote disks...
  - Inside GPU : registers, local memory, shared memory, constant memory, texture cache, processor grouping, locked physical pages, host memory access...
- Heterogeneous communications
  - Anisotropic networks
  - Various protocols

⚠ Several dimensions to cope with at the same time ☺

Outline

OpenMP

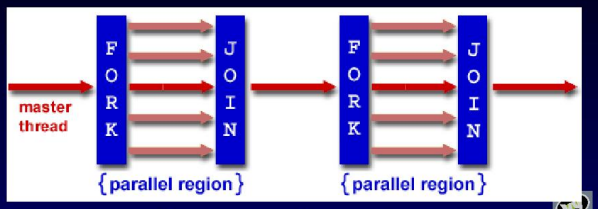
- « Le multithread pour les nuls » ☺
- Vise machines à mémoire partagée
- Sauf si programmation système compliquée, pas besoin de faire de la programmation de threads explicites
- Idée : saupoudrer un programme de directives pour aider compilateur à paralléliser
- Philosophie : `#pragma`lisme avec une certaine élégance esthétique
- ⚠ Si pas de directives, pas de parallélisme exploité (*a priori*)
- ⚠ ⚠ ⚠ Directive ≡ déclaration sur l'honneur
- Langages supportés : Fortran et C/C++

`http://openmp.org`

Modèle d'exécution d'OpenMP *(I)*

`http://openmp.org/wp/openmp-specifications`

- Exécution parallèle SPMD basée sur le *fork/join*



- Création de thread implicite ou explicite avec des *directives*

Modèle d'exécution d'OpenMP *(II)*

- Astuce : un programme OpenMP peut être exécuté comme
  - Programme séquentiel
  - Programme parallèle

  ⇝ ↗ portabilité, coût de sortie nul ! ☺
- Threads créées dans des sections `parallel` et stoppées à la fin avec une barrière
- Constructions de synchronisation et dans bibliothèque
- Contrôle de l'environnement d'exécution par variables d'environnement et fonctions de bibliothèque

Exemple

```
#pragma omp parallel default(none) \
            shared(n,x,y) private(i)
{
 /* Ceci s'exécute sur plusieurs threads en // */
#pragma omp for
  for (i=0; i<n; i++)
      // Les itérations sont réparties sur les threads
      x[i] += y[i];
 // Synchronisation implicite ici
} /*-- End of parallel region --*/
 // Synchronisation implicite ici
```


Task en OpenMP 3.0

- Rajout de la notion de tâches explicites ≡ bout de programme exécuté sur une thread
- Tâche créée dans thread par construction `task` (TBB, Cilk...)

```
#pragma omp parallel
{
  #pragma omp single private(p)
  {
   p = listhead;
   while (p) {
      #pragma omp task
      {
            process (p);
      }
      p=next(p);
   }
  }
}
```

- Extensions en vue (`target`...)

Software environments ▸ Multicores

Multimedia SIMD extensions (I)

- Multimedia and telecommunication market ↗↗
  - GIF image: 8-bit pixels
  - RGBα true-color image: 4 × 8 bit pixels
  - Phone A- or μ-law : 8-bit samples
  - CD-quality sound : 2 × 16-bit samples
- In general purpose processors, under-used transistors on these applications (double precision multiplier...) ☹
- Idea : 128-bit data viewed as independent vectors of 16 independent 8-bit elements or 8 16-bit or... 2 DP-float or 4 SP-float
- ⇝ Add SIMD to general purpose processors (i860, SSE 4.2 Core i7, AMD 3Dnow!, VMX Power, ARM, SPARC...) and Cell for numerical computing, strings, data compaction, cryptography...
- SSE4.2 at 3.2 GHz : 2 128-bit operations/cycle ⇝ 819 GOP1b/s, 102 GOP8b/s *per core*! ☺

Software environments ▸ Multicores

Multimedia SIMD extensions (II)

- Programming
  - Quite complex ☹ because many *ad hoc* instructions, saturation arithmetic...
  - Just do it... in assembly language ☹
  - Intrinsics functions in C/C++ (GCC, Intel...)
  - C/C++ extensions with new vector data types (GCC, Intel...)
  - Auto-vectorizing (IBM `xlc`, Intel, GCC, generic tools such as PIPS)
  - Use already optimized libraries
- Available on all machines ☺

Software environments ▸ Multicores

Message Passing Interface (MPI) (I)

- Le passage de message pour les nuls ! ☺
- Bibliothèque de fonctions de communication disponible pour de nombreux langages et systèmes d'exploitation
- Portabilité et nivellement par le bas : programmation de SMP aussi en MPI...
- Ressources
  - `http://en.wikipedia.org/wiki/Message_Passing_Interface`
  - MPI Forum `http://www.mpi-forum.org`
  - MPICH : A Portable Implementation of MPI `http://www-unix.mcs.anl.gov/mpi/mpich/`
  - LAM / MPI Parallel Computing `http://www.mpi.nd.edu/lam/`
  - MPE Graphics–Scalable X11 Graphics in MPI `http://www-fp.mcs.anl.gov/~lusk/papers/mpe/`
  - Livre « Using MPI: Portable Parallel Programming with the Message-Passing Interface » `http://www-unix.mcs.anl.gov/mpi/usingmpi/`

Software environments ▸ Multicores

Message Passing Interface (MPI) (II)

- Faute de temps je n'utilise plus mes transparents trop complets `http://enstb.org/~keryell/cours/MR2/IAHP/MPI`
- J'utilise « An Introduction to MPI Parallel Programming with the Message Passing Interface » de William Gropp & Ewing Lusk `http://www-unix.mcs.anl.gov/mpi/tutorial/mpiintro/MPIIntro.PPT`

Software environments ▸ Multicores

Thread Building Blocks (TBB) (I)

`http://en.wikipedia.org/wiki/TBB` from Intel

- Template library (*à la* STL)
- Open and commercial Versions
- Algorithms (`for`, `reduce`, `pipeline`, `scan`...), containers, memory allocators, mutual exclusion, atomic operations, schedulers, profiling...
- Work stealing between tasks
- Orthogonal to OpenMP & MPI

Software environments ▸ Multicores

Thread Building Blocks (TBB) (II)

```
1  class ApplyFoo {
2    float* const my_a;
   public:
4    ApplyFoo(float* a) : my_a(a) {};
     void operator() (const tbb::blocked_range<size_t>&r) const {
6      for (size_t i=r.begin(); i != r.end(); ++i)
         Foo (my_a[i]);
8    }
   }
10
   void ParallelApplyFoo(float a[], size_t n) {
12   tbb::parallel_for (
       tbb::blocked_range<size_t>(0,n),
14     ApplyFoo(a),
       tbb::auto_partitioner()
16   );
   }
```

- Need a deep code restructuring if an application is not in a STL spirit

Software environments ▶ GPU

Outline

Software environments ▶ GPU

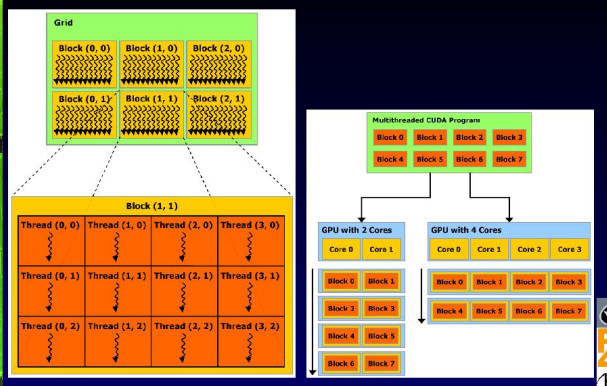
Basic GPU programming model

A sequential program on a host launches computational-intensive kernels on a GPU

- Allocate storage on the GPU
- Copy-in data from the host to the GPU
- Launch the kernel on the GPU
- The host waits...
- Copy-out the results from the GPU to the host
- Deallocate the storage on the GPU

Software environments ▶ GPU

GPU execution model



Software environments ▶ GPU

From hardware constraints to programming style (I)

- GPU computes fast but connected to CPU with slow PCI link ⇝
  - Avoid exchanging too much data between CPU & GPU or... compute on the CPU ☹
  - Possible to overlap communications with computations (more complex programing)
- Many SIMD engines (multiprocessors) ⇝ at least as much blocks of threads
- Memory hierarchy is quite complex and... visible!
  - Use (quite limited ☹) local registers by recycling local data
  - Memory is accessed in huge lines ⇝ program to use all the elements of the line
  - If not possible, try to reorganize data in the shared memory around read/write (matrix transposition...)
  - Recently added caches help too

Software environments ▶ GPU

From hardware constraints to programming style (II)

- Memory is far far away (800+ cycles) ⇝ use a lots of thread per block (but limited resources reduce block numbers) to overlap memory access with other computations
- Computing is fast, memory is slow. Rethink algorithms...
- SIMD machine, only one control flow ⇝ predicated

```
1 if (cond[i])
    b[i] = a[i];
3 else
    b[i] = -a[i] + 1;
```

- Some hardware optimizations if in a SIMD warp there is no execution ⇝ if possible sort `false`/`true` elements

Software environments ▶ GPU

Programmation CUDA (I)

- Data-parallel extension to a C++ subset
- Target nVidia GPU and *x*86 multicores
- 2-level parallelism: threads in blocks of threads + block-tiling
- In a block of threads : communication through shared memory and synchronization via `__syncthreads()`
- Complex heterogeneous memory layout (GPU...)

```
__global__ void
add_matrix_gpu(float *a, float *b, float *c, int N) {
  int i=blockIdx.x*blockDim.x+threadIdx.x;
  int j=blockIdx.y*blockDim.y+threadIdx.y;
  int index =i+j*N;

  if( i < N && j < N)
    c[index]=a[index]+b[index];
}

void main() {
  float ha[N][N], hb[N][N], hc[N][N];
```

Software environments ▶ GPU

Programmation CUDA (II)

```
/* Allocate array on the GPU with cudaMalloc */
float *a, *b, *c;
cudaMalloc((void **) &a, sizeof(float)*N*N);
cudaMalloc((void **) &b, sizeof(float)*N*N);
cudaMalloc((void **) &c, sizeof(float)*N*N);

cudaMemcpy(a, ha, sizeof(float)*N*N, cudaMemcpyHostToDevice);
cudaMemcpy(b, hb, sizeof(float)*N*N, cudaMemcpyHostToDevice);

// Describe iteration tiling (2D strip-mining)
dim3 dimBlock (blocksize,blocksize);
dim3 dimGrid (N/dimBlock.x,N/dimBlock.y);
add_matrix_gpu<<<dimGrid,dimBlock>>>(a,b,c,N);
cudaMemcpy(c, hc, sizeof(float)*N*N, cudaMemcpyDeviceToHost);
}
```

- Need some heavy code restructuring
- ∃ other version: CUDA driver, similar to OpenCL

Programming environments for big data processing on modern parallel architectures
HPC Project Ronan KERYELL 60 / 101

Software environments ▶ GPU

OpenCL (I)

- Language based on a C_{99} subset
- Started by Apple to *unify* parallel use (multicores, GPGPU...)
  ⇝ similar to OpenGL & OpenAL
- Followed by AMD/ATI and nVidia
- Data-parallelism and control-parallelism (1–3-dimensions) according to targets
- Kernel oriented computations on streams
- Complex split memory model (GPGPU...) but CPU compliant too
- New types (vectors, images...)

Programming environments for big data processing on modern parallel architectures
HPC Project Ronan KERYELL 61 / 101

Software environments ▶ GPU

OpenCL (II)

```
/* This kernel computes FFT of length 1024.
   The 1024 length FFT is decomposed into calls to a radix 16
   function, another radix 16 function and then a radix 4 function */

__kernel void fft1D_1024 (__global float2 *in, __global float2 *out,
                          __local float *sMemx, __local float *sMemy) {
  int tid = get_local_id(0);
  int blockIdx = get_group_id(0) * 1024 + tid;
  float2 data[16];
  // starting index of data to/from global memory
  in = in + blockIdx;  out = out + blockIdx;
  globalLoads(data, in, 64); // coalesced global reads
  fftRadix16Pass(data);      // in-place radix-16 pass
  twiddleFactorMul(data, tid, 1024, 0);
  // local shuffle using local memory
  localShuffle(data, sMemx, sMemy, tid,
               (((tid & 15) * 65) + (tid >> 4)));
  fftRadix16Pass(data);               // in-place radix-16 pass
  twiddleFactorMul(data, tid, 64, 4); // twiddle factor multiplication
  localShuffle(data, sMemx, sMemy, tid,
               (((tid >> 4) * 64) + (tid & 15)));
  // four radix-4 function calls
  fftRadix4Pass(data); fftRadix4Pass(data + 4);
  fftRadix4Pass(data + 8); fftRadix4Pass(data + 12);
```

Programming environments for big data processing on modern parallel architectures
HPC Project Ronan KERYELL 62 / 101

Software environments ▶ GPU

OpenCL (III)

```
  // coalesced global writes
  globalStores(data, out, 64);
}
[...]
// create a compute context with GPU device
context = clCreateContextFromType(CL_DEVICE_TYPE_GPU);
// create a work-queue
queue = clCreateWorkQueue(context, NULL, NULL, 0);
// allocate the buffer memory objects
memobjs[0] = clCreateBuffer(context, CL_MEM_READ_ONLY |
CL_MEM_COPY_HOST_PTR,
                            sizeof(float)*2*num_entries, srcA);
memobjs[1] = clCreateBuffer(context, CL_MEM_READ_WRITE,
                            sizeof(float)*2*num_entries, NULL);
// create the compute program
program = clCreateProgramFromSource(context, 1,
                                    &fft1D_1024_kernel_src, NULL);
// build the compute program executable
clBuildProgramExecutable(program, false, NULL, NULL);
// create the compute kernel
kernel = clCreateKernel(program, "fft1D_1024");
// create N-D range object with work-item dimensions
global_work_size[0] = n;
local_work_size[0] = 64;
```

Programming environments for big data processing on modern parallel architectures
HPC Project Ronan KERYELL 63 / 101

Software environments ▶ GPU

OpenCL (IV)

```
range = clCreateNDRangeContainer(context, 0, 1,
                                 global_work_size, local_work_size);
// set the args values
clSetKernelArg(kernel, 0, (void *)&memobjs[0], sizeof(cl_mem), NULL);
clSetKernelArg(kernel, 1, (void *)&memobjs[1], sizeof(cl_mem), NULL);
clSetKernelArg(kernel, 2, NULL,
               sizeof(float)*(local_work_size[0]+1)*16, NULL);
clSetKernelArg(kernel, 3, NULL,
               sizeof(float)*(local_work_size[0]+1)*16, NULL);
 // execute kernel
clExecuteKernel(queue, kernel, NULL, range, NULL, 0, NULL);
```

- Need a lot of code restructuring

Programming environments for big data processing on modern parallel architectures
HPC Project Ronan KERYELL 64 / 101

Software environments ▶ GPU

CUDA or OpenCL?

CUDA

- Appeared first
- Language basis not well defined: C++ like, rather C89 and not C99
- Painful to translate C99 to C89 and keeping clean sources
- Rather limited to 2D threads
- nVidia GPU only

OpenCL

- Standard backed by many companies
- C99 based ⇝ clean
- 3D threads with less constraints
- More verbose API (kernel call...)
- Kernel source code outside of host source: more complex
- Fast spreading in embedded computing world (MP-SoC)

Programming environments for big data processing on modern parallel architectures
HPC Project Ronan KERYELL 65 / 101

Software environments ▸ Application libraries

Outline

Software environments ▸ Application libraries

Bibliothèques mathématiques (I)

- Existent pour diffrentes architectures !
- Souvent bibliothèques constructeurs optimisées pour leur machines
  - Intel Math Kernel Library (MKL)
  - AMD Performance Library (⇝ Framewave libre)
- FFT : FFTW (*Fastest Fourier Transform in the West*, en C généré par du OCaml)...
- Beaucoup d'algèbre linéaire
  - BLAS (*Basic Linear Algebra Subprograms*) et PBLAS
  - LAPACK (*Linear Algebra PACKage*)
  - ScaLAPACK : version SPMD avec MPI
  - SuperLU : solution directe de gros systèmes creux
  - PETSc (*Portable, Extensible Toolkit for Scientific Computation*) : large spectre, au dessus de MPI

Par4All

Outline

Par4All

Use the Source, Luke...

Hardware is moving quite (too) fast but...

What has survived for 50+ years?

Fortran programs...

What has survived for 40+ years?

IDL, Matlab, Scilab...

What has survived for 30+ years?

C programs, Unix...

- A lot of legacy code could be pushed onto parallel hardware (accelerators) with automatic tools...
- Need automatic tools for source-to-source transformation to leverage existing software tools for a given hardware
- Not as efficient as hand-tuned programs, but quick production phase

Par4All

Not reinventing the wheel... No NIH syndrome please!

Want to create your own tool?

- House-keeping and infrastructure in a compiler is a huge task
- Unreasonable to begin yet another new compiler project...
- Many academic Open Source projects are available...
- ...But customers need products ☺
- ⇝ Integrate your ideas and developments in existing project
- ...or buy one if you can afford (ST with PGI...) ☺
- Some projects to consider
  - Old projects: gcc, PIPS... and many dead ones (SUIF...)
  - But new ones appear too: LLVM, RoseCompiler, Cetus...

Par4All

- ⇝ Funding an initiative to industrialize Open Source tools
- Use source-to-source tool to be more target-independent
- PIPS is the first project to enter the Par4All initiative

Par4All

Current PIPS usage

Developed for 23 years (...) @ Mines ParisTech & Télécom Bretagne, mainly

- Automatic parallelization (Par4All C & Fortran to OpenMP)
- Distributed memory computing with OpenMP-to-MPI translation [STEP project]
- Generic vectorization for SIMD instructions (SSE, VMX, Neon, CUDA, OpenCL...) (SAC project) [SCALOPES]
- Parallelization for embedded systems [SCALOPES]
- Compilation for hardware accelerators (Ter@PIX, SPoC, SIMD, FPGA, MPPA, P2012...) [FREIA, SCALOPES, SMECY]
- High-level hardware accelerators synthesis generation for FPGA [PHRASE, CoMap]
- Reverse engineering & decompiler (reconstruction from binary to C)
- Genetic algorithm-based optimization [Luxembourg university+TB]
- Code instrumentation for performance measures
- GPU with CUDA & OpenCL [TransMedi@, FREIA, OpenGPU]

Par4All usage

- Generate from sequential C, Fortran & Scilab code
  - OpenMP for SMP
  - CUDA for nVidia GPU
  - OpenCL for GPU & ST Platform 2012 (on-going)
  - Code for various accelerators [SMECY], Kalray [SIMILAN]... (on-going)
  - SCMP task programs for SCMP machine from CEA and for... cloud computing ☺

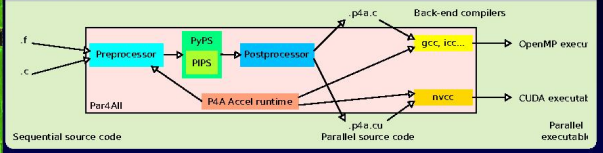
Par4All ≡ PyPS scripting in the backstage (I)

- PIPS is a great tool-box to do source-to-source compilation
- ...but not really usable by λ end-user ☹
- ⇝ Development of Par4All
- Add a user compiler-like infrastructure
- ⇝ `p4a` script as simple as
  - `p4a --openmp toto.c -o toto`
  - `p4a --cuda toto.c -o toto -lm`
- Be multi-target
- Apply some adaptative transformations
- Up to now PIPS was scripted with a special shell-like language: `tpips`
- Not enough powerful (not a programming language)
- Develop a SWIG Python interface to PIPS phases and interface

Par4All ≡ PyPS scripting in the backstage (II)

- All the power of a widely spread real language
- Automate with introspection through the compilation flow
- Easy to add any glue, pre-/post-processing to generate target code

Overview



- Invoke PIPS transformations
  - With different recipes according to generated stuff
  - Special treatments on kernels...
- Compilation and linking infrastructure: can use `gcc`, `icc`, `nvcc`, `nvcc+gcc`, `nvcc+icc`

Par4All ≡ PyPS scripting in the backstage (III)

- House keeping code
- Fundamental: colorizing and filtering some PIPS output, running cursor... ☺

Coding rules

- Automatic parallelization is not magic
- Use abstract interpretation to « understand » programs
- Undecidable in the generic case ($\approx$ halting problem)
- Quite easier for well written programs
- Develop a coding rule manual to help parallelization and... sequential quality!
  - Avoid useless pointers
  - Take advantage of C99 (arrays of non static size...)
  - Use higher-level C, do not linearize arrays...
  - Organize execution in cleaner loops expressing better parallelism
  - ...
- Prototype of coding rules report on-line on `par4all.org`

Parallelization to OpenMP (I)

- The easy way... Already in PIPS
- Used to bootstrap the start-up with stage-0 investors ☺
- Indeed, we used only `bash`-generated `tpips` at this time (2008, no PyPS yet), but needed a lot of bug squashing on C support in PIPS...

OpenMP output sample

```
!$omp parallel do private(I, K, X)
C multiply the two square matrices of ones
      DO J = 1, N
0016
!$omp parallel do private(K, X)
         DO I = 1, N
0017
            X = 0
0018
!$omp parallel do reduction(+:X)
            DO K = 1, N
0019
               X = X+A(I,K)*B(K,J)
0020
            ENDDO
!$omp end parallel do
            C(I,J) = X
0022
         ENDDO
!$omp end parallel do
      ENDDO
!$omp end parallel do
```


Outline

Basic GPU execution model

A sequential program on a host launches computational-intensive kernels on a GPU

- Allocate storage on the GPU
- Copy-in data from the host to the GPU
- Launch the kernel on the GPU
- The host waits...
- Copy-out the results from the GPU to the host
- Deallocate the storage on the GPU

Generic scheme for other heterogeneous accelerators too

Scilab & Matlab

- Languages used for simulations, data analytics, pricing...
- Scilab/Matlab input : *sequential* or array syntax
- Compilation to C code
  - Side effect of MediaGPU ANR project...
  - Our COLD compiler is *not* Open Source
  - There is such Open Source compiler from hArtes European project written in... Scilab ☺
- Parallelization of the generated C code
- Use parallel runtime too
- Type inference to guess (crazy ☺) semantics
  - Heuristic: first encountered type is forever
- May get speedup ≫ 1000 ☺
- Wild Cruncher product from HPC Project: *x86+GPU* appliance with nice interface
  - Scilab — mathematical model & simulation
  - Par4All — automatic parallelization
  - //Geometry — polynomial-based 3D rendering & modelling

Hyantes (I)

- Geographical application for data integration: library to compute neighbourhood population potential with scale control
- WildNode with 2 Intel Xeon X5670 @ 2.93GHz (12 cores) and a nVidia Tesla C2050 (Fermi), Linux/Ubuntu 10.04, gcc 4.4.3, CUDA 3.1
  - Sequential execution time on CPU: 30.355s
  - OpenMP parallel execution time on CPUs: 3.859s, speed-up: 7.87
  - CUDA parallel execution time on GPU: 0.441s, speed-up: 68.8
- With single precision on a HP EliteBook 8730w laptop (with an Intel Core2 Extreme Q9300 @ 2.53GHz (4 cores) and a nVidia GPU Quadro FX 3700M (16 multiprocessors, 128 cores, architecture 1.1)) with Linux/Debian/sid, gcc 4.4.5, CUDA 3.1:
  - Sequential execution time on CPU: 34.7s
  - OpenMP parallel execution time on CPUs: 13.7s, speed-up: 2.53
  - OpenMP emulation of GPU on CPUs: 9.7s, speed-up: 3.6
  - CUDA parallel execution time on GPU: 1.57s, speed-up: 24.2

Hyantes (II)

Original main C kernel:

```
void run(data_t xmin, data_t ymin, data_t xmax, data_t ymax, data_t step, d
  town pt[rangex][rangey], town t[nb])
{
    size_t i,j,k;

    fprintf(stderr,"begin_computation_...\n");

    for(i=0;i<rangex;i++)
        for(j=0;j<rangey;j++) {
            pt[i][j].latitude =(xmin+step*i)*180/M_PI;
            pt[i][j].longitude =(ymin+step*j)*180/M_PI;
            pt[i][j].stock =0.;
            for(k=0;k<nb;k++) {
                data_t tmp = 6368.* acos(cos(xmin+step*i)*cos( t[k].latitud
                    * cos((ymin+step*j)-t[k].longitude)
                    + sin(xmin+step*i)*sin(t[k].latitude));
                if( tmp < range )
                    pt[i][j].stock += t[k].stock  / (1 + tmp) ;
            }
        }
    fprintf(stderr,"end_computation_...\n");
}
```

Par4All ▶ GPU code generation

Hyantes (III)

Example given in `par4all.org` distribution

Par4All ▶ GPU code generation

Hyantes (IV)

OpenMP code:

```
void run(data_t xmin, data_t ymin, data_t xmax, data_t ymax, data_t step, d
{
   size_t i, j, k;

   fprintf(stderr, "begin_computation_...\n");

#pragma omp parallel for private(k, j)
   for(i = 0; i <= 289; i += 1)
      for(j = 0; j <= 298; j += 1) {
         pt[i][j].latitude = (xmin+step*i)*180/3.14159265358979323846;
         pt[i][j].longitude = (ymin+step*j)*180/3.14159265358979323846;
         pt[i][j].stock = 0.;
         for(k = 0; k <= 2877; k += 1) {
            data_t tmp = 6368.*acos(cos(xmin+step*i)*cos(t[k].latitude)*cos
            if (tmp<range)
               pt[i][j].stock += t[k].stock/(1+tmp);
         }
      }
   fprintf(stderr, "end_computation_...\n");
}
void display(town pt[290][299])
{
```


Par4All ▶ GPU code generation

Hyantes (V)

```
   size_t i, j;
   for(i = 0; i <= 289; i += 1) {
      for(j = 0; j <= 298; j += 1)
         printf("%lf_%lf_%lf\n", pt[i][j].latitude, pt[i][j].longitude, pt[
      printf("\n");
   }
}
```


Par4All ▶ GPU code generation

Hyantes (VI)

Generated GPU code:

```
void run(data_t xmin, data_t ymin, data_t xmax, data_t ymax, data_t step, d
  town pt[290][299], town t[2878])
{
   size_t i, j, k;
   //PIPS generated variable
   town (*P_0)[2878] = (town (*)[2878]) 0, (*P_1)[290][299] = (town (*)[290

   fprintf(stderr, "begin_computation_...\n");
   P4A_accel_malloc(&P_1, sizeof(town[290][299])-1+1);
   P4A_accel_malloc(&P_0, sizeof(town[2878])-1+1);
   P4A_copy_to_accel(pt, *P_1, sizeof(town[290][299])-1+1);
   P4A_copy_to_accel(t, *P_0, sizeof(town[2878])-1+1);

   p4a_kernel_launcher_0(*P_1, range, step, *P_0, xmin, ymin);
   P4A_copy_from_accel(pt, *P_1, sizeof(town[290][299])-1+1);
   P4A_accel_free(*P_1);
   P4A_accel_free(*P_0);
   fprintf(stderr, "end_computation_...\n");
}

void p4a_kernel_launcher_0(town pt[290][299], data_t range, data_t ste
  data_t xmin, data_t ymin)
```


Par4All ▶ GPU code generation

Hyantes (VII)

```
{
   //PIPS generated variable
   size_t i, j, k;
   P4A_call_accel_kernel_2d(p4a_kernel_wrapper_0, 290,299, i, j, pt, range,
                            step, t, xmin, ymin);
}

P4A_accel_kernel_wrapper void p4a_kernel_wrapper_0(size_t i, size_t j, town
  data_t range, data_t step, town t[2878], data_t xmin, data_t ymin)
{
   // Index has been replaced by P4A_vp_0:
   i = P4A_vp_0;
   // Index has been replaced by P4A_vp_1:
   j = P4A_vp_1;
   // Loop nest P4A end
   p4a_kernel_0(i, j, &pt[0][0], range, step, &t[0], xmin, ymin);
}

P4A_accel_kernel void p4a_kernel_0(size_t i, size_t j, town *pt, data_t ran
  data_t step, town *t, data_t xmin, data_t ymin)
{
   //PIPS generated variable
   size_t k;
   // Loop nest P4A end
```


Par4All ▶ GPU code generation

Hyantes (VIII)

```
   if (i<=289&&j<=298) {
      pt[299*i+j].latitude = (xmin+step*i)*180/3.14159265358979323846;
      pt[299*i+j].longitude = (ymin+step*j)*180/3.14159265358979323846;
      pt[299*i+j].stock = 0.;
      for(k = 0; k <= 2877; k += 1) {
         data_t tmp = 6368.*acos(cos(xmin+step*i)*cos((*(t+k)).latitude)*co
                    -(*(t+k)).longitude)+sin(xmin+step*i)*sin((*(t+k)).latitude)
         if (tmp<range)
            pt[299*i+j].stock += t[k].stock/(1+tmp);
      }
   }
}
```

Stars-PM

- *Particle-Mesh* N-body cosmological simulation
- C code from Observatoire Astronomique de Strasbourg
- Use FFT 3D
- Example given in `par4all.org` distribution

Stars-PM time step (I)

```
void iteration(coord pos[NP][NP][NP],
               coord vel[NP][NP][NP],
               float dens[NP][NP][NP],
               int data[NP][NP][NP],
               int histo[NP][NP][NP]) {
  /* Split space into regular 3D grid: */
  discretisation(pos, data);
  /* Compute density on the grid: */
  histogram(data, histo);
  /* Compute attraction potential
     in Fourier's space: */
  potential(histo, dens);
  /* Compute in each dimension the resulting forces and
     integrate the acceleration to update the speeds: */
  forcex(dens, force);
  updatevel(vel, force, data, 0, dt);
  forcey(dens, force);
  updatevel(vel, force, data, 1, dt);
  forcez(dens, force);
  updatevel(vel, force, data, 2, dt);
  /* Move the particles: */
  updatepos(pos, vel);
}
```


Stars-PM & Jacobi results with `p4a 1.1.2`

- 2 Xeon Nehalem X5670 (12 cores @ 2,93 GHz)
- 1 GPU nVidia Tesla C2050 CUDA 3.2
- Automatic call to CuFFT instead of FFTW (stubs...)
- 150 iterations of Stars-PM

Execution time	`p4a`	Simulation Cosmo. 32^3	64^3	128^3	Jacobi
Sequential	`(gcc -O3)`	0,68	6,30	98,4	24,5
OpenMP *6 threads*	`--openmp`	0,16	1,28	16,6	13,8
CUDA base	`--cuda`	0,88	5,21	31,4	67,7
Optim. comm. 1.1	`--cuda --com-opt.`	0,20	1,17	8,9	6,5
Reduction Optim. 1.1.2	`--cuda --com-opt.`	0,10	0,32	2,1	3,8
Manual optim.	`(gcc -O3)`	0,05	0,26	1,8	

`p4a 1.1.2` introduce generation of CUDA atomic updates for PIPS detected reductions. Other solution to investigate: CuDPP call generation

Outline

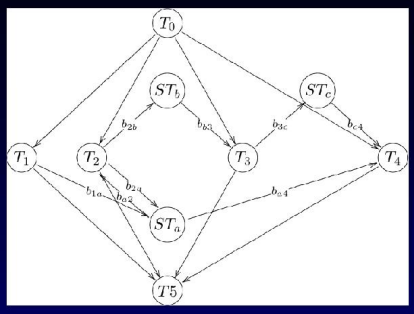
SCMP computer (I)

Future revolutions

Software radio, cognitive radio, passive radar, compressed sensing...

- Embedded accelerator developed at French CEA
  - Task graph oriented parallel multiprocessor
  - Hardware task graph scheduler
  - Synchronizations
  - Communication through memory page sharing
- Generating code from THALES (TCF) GSM sensing application in SCALOPES European project
- Reuse output of PIPS GPU phases + specific phases
  - SCMP code with tasks
  - SCMP task descriptor files
- Adapted Par4All Accel run-time

SCMP tasks



In general case, different tasks can produce data in unpredictable way: use helper data server tasks to deal with coherency when several producers

Par4All ▶ Code generation for SCMP

SCMP task code (before/after)

```
int main() {
  int i, t, a[20], b[20];
  for (t=0; t < 100; t++)
  {
kernel_tasks_1:
    for(i=0; i <10; i++)
      a[i] = i+t;

kernel_tasks_2:
    for(i=10; i <20; i++)
      a[i] = 2*i+t;

kernel_tasks_3:
    for(i=10; i <20; i++)
      printf("a[%d]_=_%d\n",
             i, a[i]);
  }
  return(0);
}
```

```
int main() {
 P4A_scmp_reset();
 int i, t, a[20], b[20];
 for(t = 0; t <= 99; t += 1) {
  [...]
  {
   //PIPS generated variable
   int (*P4A__a__1)[10] = (int (*)[10]) 0;
   P4A_scmp_malloc((void **) &P4A__a__1,
     sizeof(int)*10, P4A__a__1_id,
     P4A__a__1_prod_p || P4A__a__1_cons_p, P4A__
   if (scmp_task_2_p)
     for(i = 10; i <= 19; i += 1)
       (*P4A__a__1)[i-10] = 2*i+t;
     P4A_copy_from_accel_1d(sizeof(int), 20, 10,
       P4A_sesam_server_a_p ? &a[0] : NULL, *P4A
       P4A__a__1_id, P4A__a__1_prod_p || P4A__a_
     P4A_scmp_dealloc(P4A__a__1, P4A__a__1_id,
       P4A__a__1_prod_p || P4A__a__1_cons_p, P4A
  }
  [...]
 }
 return(ev_T004);
}
```

Programming environments for big data processing on modern parallel architectures
HPC Project Ronan KERYELL 96 / 101

Par4All ▶ Code generation for SCMP

Performance of GSM sensing on SCMP

- Speed-up on 4 PE SCMP:
  - ×2.35 with manual parallelization by SCMP team
  - ×1.86 with automatic Par4All parallelization
- Still big memory overhead
- To optimize...
- Use these techniques to generate automate cloud-ification ☺

Programming environments for big data processing on modern parallel architectures
HPC Project Ronan KERYELL 97 / 101

Par4All ▶ Code generation for SCMP

Performance of GSM sensing on SCMP

- Speed-up on 4 PE SCMP:
  - ×2.35 with manual parallelization by SCMP team
  - ×1.86 with automatic Par4All parallelization
- Still big memory overhead
- To optimize...
- Use these techniques to generate automate cloud-ification ☺

Programming environments for big data processing on modern parallel architectures
HPC Project Ronan KERYELL 97 / 101

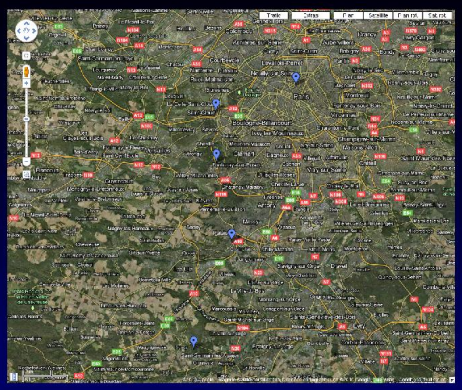
Conclusion

Outline

Programming environments for big data processing on modern parallel architectures
HPC Project Ronan KERYELL 98 / 101

Conclusion

French advantage: Saint (*Holly*) Cloud localization

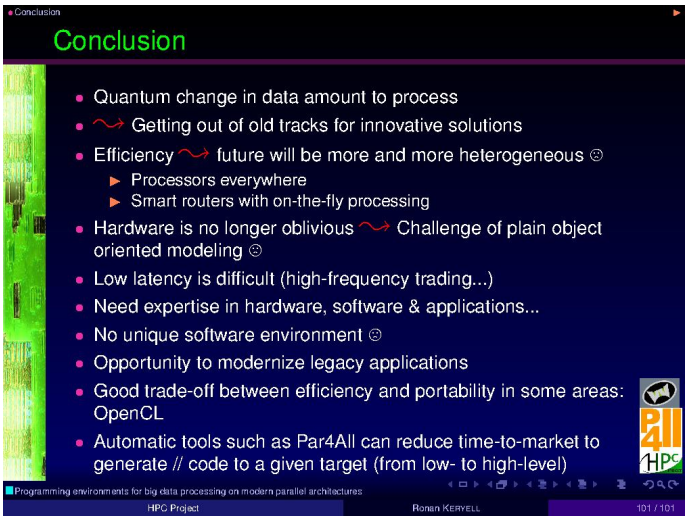


Programming environments for big data processing on modern parallel architectures
HPC Project Ronan KERYELL 99 / 101

Conclusion

Saint Cloud gatekeeper & massive virtual I/O

Programming environments for big data processing on modern parallel architectures
HPC Project Ronan KERYELL 100 / 101


Conclusion
Conclusion
Quantum change in data amount to process
⇝ Getting out of old tracks for innovative solutions
Efficiency ⇝ future will be more and more heterogeneous ☹
Processors everywhere
Smart routers with on-the-fly processing
Hardware is no longer oblivious ⇝ Challenge of plain object oriented modeling ☺
Low latency is difficult (high-frequency trading...)
Need expertise in hardware, software & applications...
No unique software environment ☹
Opportunity to modernize legacy applications
Good trade-off between efficiency and portability in some areas: OpenCL
Automatic tools such as Par4All can reduce time-to-market to generate // code to a given target (from low- to high-level)
Programming environments for big data processing on modern parallel architectures
HPC Project
Ronan KERYELL
101 / 101

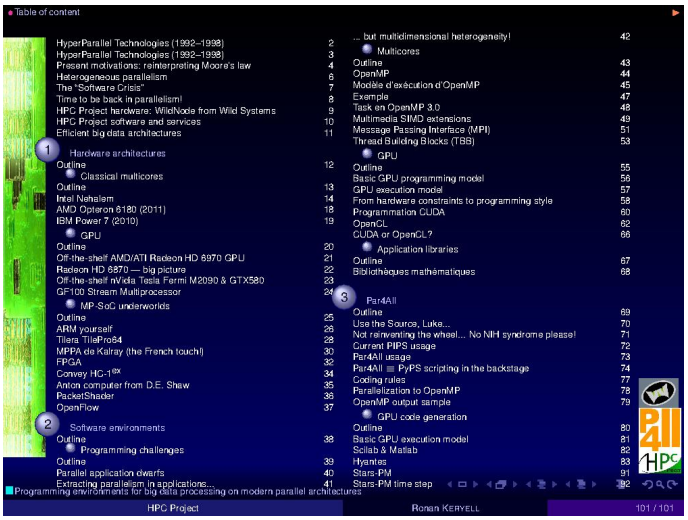


Table of content
HyperParallel Technologies (1992–1998) 2
HyperParallel Technologies (1992–1998) 3
Present motivations: reinterpreting Moore's law 4
Heterogeneous parallelism 6
The "Software Crisis" 7
Time to be back in parallelism! 8
HPC Project hardware: WildNode from Wild Systems 9
HPC Project software and services 10
Efficient big data architectures 11
1 Hardware architectures
Outline 12
Classical multicores
Outline 13
Intel Nehalem 14
AMD Opteron 8180 (2011) 18
IBM Power 7 (2010) 19
GPU
Outline 20
Off-the-shelf AMD/ATI Radeon HD 6970 GPU 21
Radeon HD 6870 — big picture 22
Off-the-shelf nVidia Tesla Fermi M2090 & GTX580 23
GF100 Stream Multiprocessor 24
MP-SoC underworlds
Outline 25
ARM yourself 26
Tilera TilePro64 28
MPPA de Kalray (the French touch!) 30
FPGA 32
Convey HC-1ex 34
Anton computer from D.E. Shaw 35
PacketShader 36
OpenFlow 37
2 Software environments
Outline 38
Programming challenges
Outline 39
Parallel application dwarfs 40
Extracting parallelism in applications... 41
... but multidimensional heterogeneity! 42
Multicores
Outline 43
OpenMP 44
Modèle d'exécution d'OpenMP 45
Exemple 47
Task en OpenMP 3.0 48
Multimedia SIMD extensions 49
Message Passing Interface (MPI) 51
Thread Building Blocks (TBB) 53
GPU
Outline 55
Basic GPU programming model 56
GPU execution model 57
From hardware constraints to programming style 58
Programmation CUDA 60
OpenCL 62
CUDA or OpenCL? 66
Application libraries
Outline 67
Bibliothèques mathématiques 68
3 Par4All
Outline 69
Use the Source, Luke... 70
Not reinventing the wheel... No NIH syndrome please! 71
Current PIPS usage 72
Par4All usage 73
Par4All ≅ PyPS scripting in the backstage 74
Coding rules 77
Parallelization to OpenMP 78
OpenMP output sample 79
GPU code generation
Outline 80
Basic GPU execution model 81
Scilab & Matlab 82
Hyantes 83
Stars-PM 91
Stars-PM time step 92
Programming environments for big data processing on modern parallel architectures
HPC Project
Ronan KERYELL
101 / 101

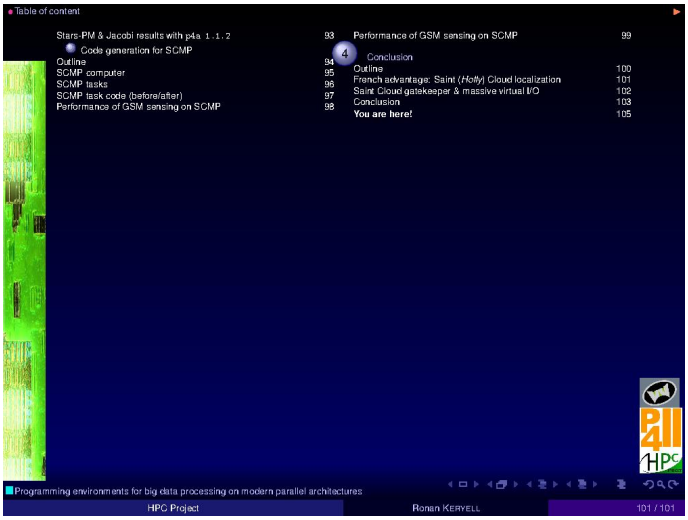


Table of content
Stars-PM & Jacobi results with p4a 1.1.2 93
Code generation for SCMP
Outline 94
SCMP computer 95
SCMP tasks 96
SCMP task code (before/after) 97
Performance of GSM sensing on SCMP 98
Performance of GSM sensing on SCMP 99
4 Conclusion
Outline 100
French advantage: Saint (Holly) Cloud localization 101
Saint Cloud gatekeeper & massive virtual I/O 102
Conclusion 103
You are here! 105
Programming environments for big data processing on modern parallel architectures
HPC Project
Ronan KERYELL
101 / 101

2.6 Denis Caromel (ActiveEon-INRIA)

Solutions `ProActive` pour *Workflows*, *Map/Reduce*, `Matlab/Scilab`, `CPU/GPU`

`ProActive Parallel Suite` (`http://ProActive.inria.fr`), un projet *Open Source* d'`OW2`, offre une solution flexible pour regrouper des ressources de calcul et offrir aux entreprises un accès simple et unifié à ces ressources par le biais de Portails et d'API. `ProActive` optimise l'exécution des applications les plus exigeantes, les *workflows* d'entreprises, les simulations numériques et financières, l'analyse des données (avec un *Map/Reduce* qui supporte les APIs `Hadoop` sans nécessiter un *cluster* dédié). Un mécanisme de sélection de ressources permet de combiner les exécutions sur `CPU` et `GPUs`, sur des tâches simples ou au sein même de *workflows* dynamiques. Des analyses de donnés parallèles sur *Cloud* peuvent être initiées direcrement sans quitter les environnements `Matlab` et `Scilab`. Des *benchmarks* dans les *biotechs* et des démonstrations interactives de *Map/Reduce* seront présentées sur une plate-forme en production.

ProActive Parallel Suite

Solutions ProActive pour Workflows Map/Reduce, Matlab/Scilab, CPU/GPU

Denis Caromel (INRIA & ActiveEon)
Cédric Dalmasso (ActiveEon)

Accelerate and Orchestrate Enterprise Applications

Hybrid Cloud Solutions (CPU+GPU, Private, Public Burst, Multi-Tenants)

Le déluge de données, Ecole Polytechnique, Palaiseau, June 9th 2011

ProActive Parallel Suite

INRIA OASIS Team Composition (35)

- Researchers (5):
  - D. Caromel (UNSA, Det. INRIA
  - E. Madelaine (INRIA)
  - F. Baude (UNSA)
  - F. Huet (UNSA)
  - L. Henrio (CNRS)
- PhDs (11):
  - Antonio Cansado (INRIA, Coni
  - Brian Amedro (SCS-Agos)
  - Cristian Ruz (INRIA, Conicyt)
  - Elton Mathias (INRIA-Cordi)
  - Imen Filali (SCS-Agos / FP7 S
  - Marcela Rivera (INRIA, Conicy
  - Muhammad Khan (STIC-Asia)
  - Paul Naoumenko (INRIA/Régio
  - Viet Dung Doan (FP6 Bionets)
  - Virginie Contes (SOA4ALL)
  - Guilherme Pezzi (AGOS, CIFR
- + Visitors + Interns

Located in Sophia Antipolis, between Nice and Cannes,
Visitors Welcome!

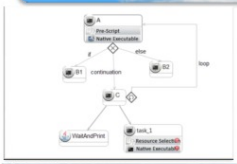
ActiveEon Overview

- **ActiveEon**, a software company **born of INRIA**, founded in 2007HQ in the French scientific park Sophia Antipolis
- **Co developing** with INRIA ***ProActive Parallel Suite®***, a Professional Open Source middleware for parallel, distributed, multi-core computing 30 peoples in total
- Core **mission**: Scale Beyond Limits
- Providing **a full range of services** for ProActive Parallel Suite
- **Worldwide** customers and production users:

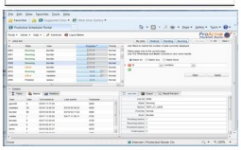
RENAULT, hp, IFP Energies nouvelles, orange, EDF, GMO, DEXIA, THALES, sirehna, CEA, AMADEUS, EMBL EBI

ProActive Parallel Suite

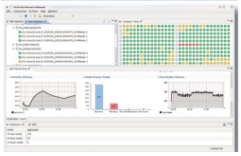
ProActive Parallel Suite



Workflow Execution
Studio Editor and Visualization

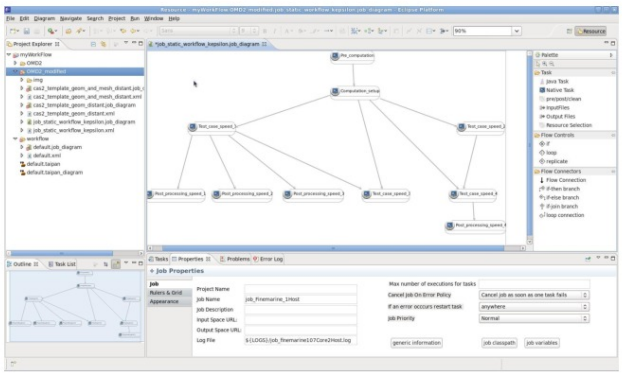


Portal, Multi-Application & Multi-Tenant Enterprise Orchestration



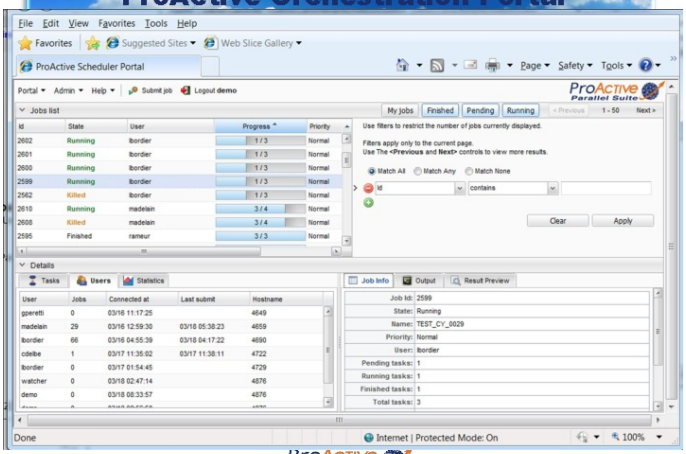
Physical and Virtual Machines Management

Workflow Studio



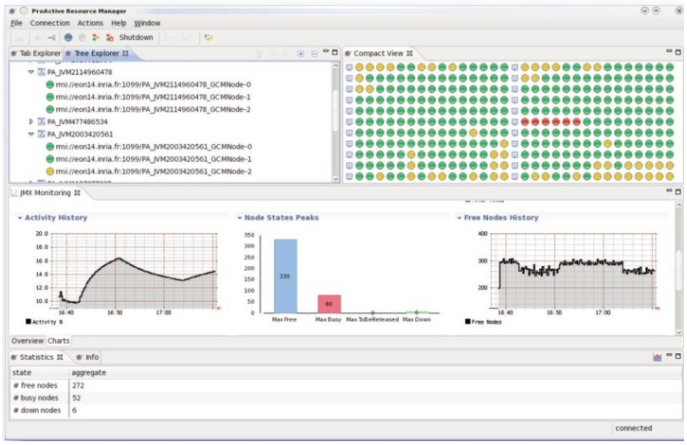
ProActive Parallel Suite

ProActive Orchestration Portal



ProActive Parallel Suite

RESOURCING User Interface



OW2 Consortium

ProActive Parallel Suite Map Reduce

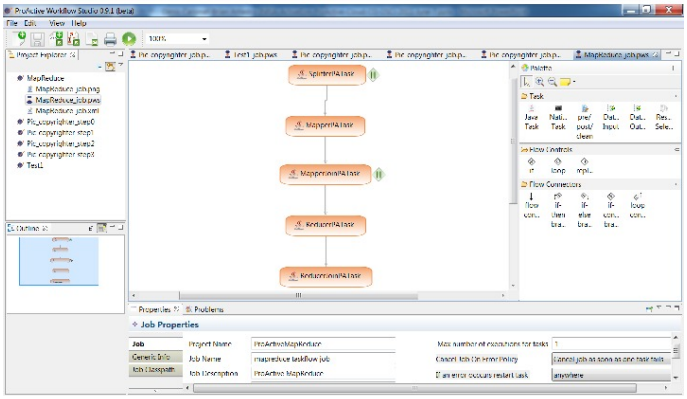
Live Demo

ProActive MapReduce (CO, SP2, Task 2.1)

- Same APIs as Hadoop
  (Easy switch from Hadoop to ProActive)
- Does not requires an HDFS File System
- Runs on general purpose, Multi-tenant, Multi-Applications Grids and Clouds
- Available as PaaS in Java

ProActive Parallel Suite

Workflow ProActive MapReduce



ProActive Parallel Suite

Map Reduce Demo

ProActive Parallel Suite

ProActive MapReduce vs. Hadoop+HDFS

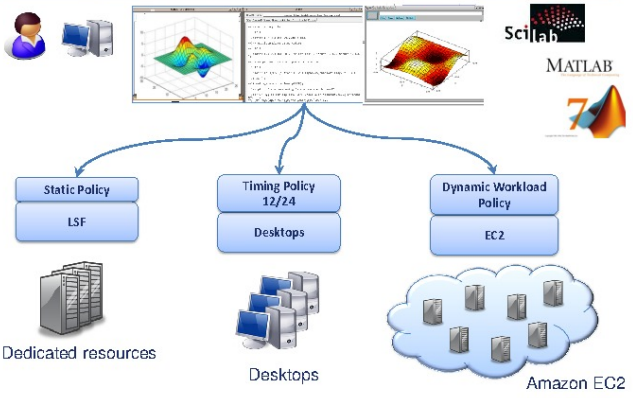
File Size	Sequential	Hadoop	PA MapReduce	Speedup
0.7 GB	5m 04s	1m 17s	1m 05s	4.6
4.3 GB	25m 31s	2m 30s	2m 20s	10.9
7.3 GB	46m 00s	3m 31s	3m 30s	13.1
20 GB	2h 07m 00s	8m 30s	7m 09s	17.8
50 GB	5h 19m 00s	21m 05s	25m 11s	12.7
100 GB	10h 38m 00s	43m 23s	58m 42s	10.9

- Data available in a NAS (General purpose storage)
- Transfer to HDFS for Hadoop
- Used directly without copy for ProActive
- Use Case of Map/Reduce on fresh data
- Different ProActive Map/Reduce configuration for recurrent MR on in place Data (e.g. ProActive HDFS interface)

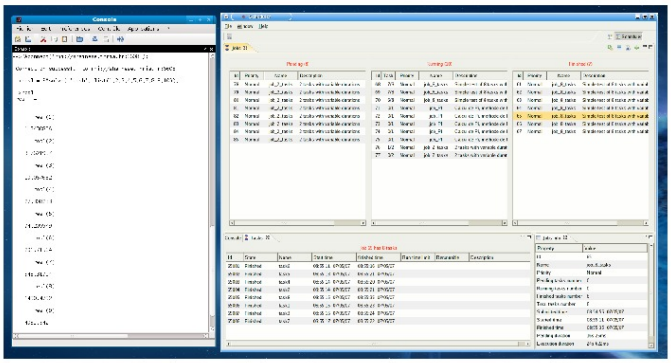
ProActive Parallel Suite

ProActive Parallel Suite With Scilab MATLAB

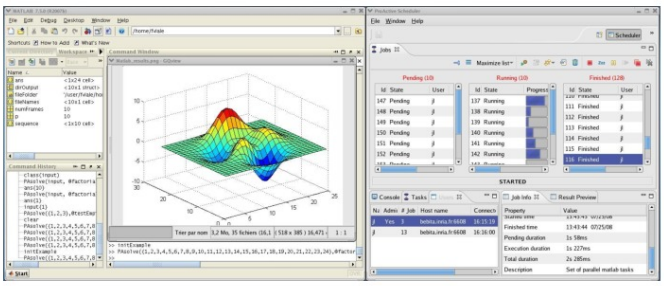
Integration with Scilab and Matlab



Interface ProActive ⇔ Scilab



Interface ProActive ⇔ Matlab



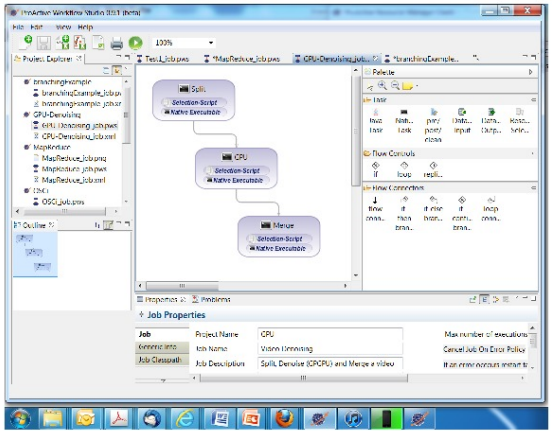
ProActive Parallel Suite CPU + GPU Workflow

Live Demo

CPU + GPU ProActive Workflows

- Resource selection for each Task of a ProActive Workflow
- Selection of Host with GPU capacity
- Data Transfer to the GPU Host
- Configuration of GPU Capacity at the level of Admin (Number of GPU Nodes, size)
- Freedom to request one or several GPU capacities for one GPU program
- Global Scheduling (Multi-Tenant, Multi-Application) of GPU Tasks

Workflow ProActive for CPU and GPU



19

GPU CPU Demo

ProActive Parallel Suite

20

Conclusion

ProActive Parallel Suite

22

Conclusion:

ProActive Parallel Suite

Workflow Studio editor
Workflow Execution (with Visualization)
Map/Reduce, Matlab, Scilab, CPU and GPU

Portal, and APIs (Java, REST, ...)
Multi-Application & Multi-Tenant
Enterprise Orchestration

Physical and Virtual Machines Management
(Hyper-V, VMware, VirtualBox, KVM, Qemu, Xen)
Public Cloud (EC2, Windows Azure), ...

ProActive Parallel Suite

22

2.7 Nicolas Pons (INRA)

La métagénomique, un défi supplémentaire pour la loi de Moore

Depuis le séquençage du premier génome en 1995, la production de données de séquençage d'ADN a révolutionné les possibilités de compréhension du vivant par la biologie moléculaire. Avec l'arrivée des technologies de séquençage à très haut-débit, on assiste aujourd'hui à une explosion des volumes de données avec un doublement des bases de données de séquence tous les 6 mois et une augmentation du débit d'acquisition d'un facteur 1000. Ce déluge de données ouvre de nouvelles perspectives scientifiques notamment dans le domaine de la métagénomique qui vise à caractériser l'ensemble des génomes bactériens d'un écosystème complexe : il est désormais possible de quantifier les génomes, gènes et fonctions de ces écosystèmes. Le traitement de ces *big data* constitue un défi majeur tant en matière d'optimisation des calculs qu'en matière de stockage et de leur mise à disposition aux biologistes. Nous illustrerons ces défis à travers l'exemple des projets `MetaHIT` et `MicroObes` qui proposent d'étudier le génome de l'ensemble des bactéries constituant la flore intestinale humaine afin de caractériser ses fonctions et ses implications sur la santé.

Aristote *Le déluge de données, comment en tirer parti ?*

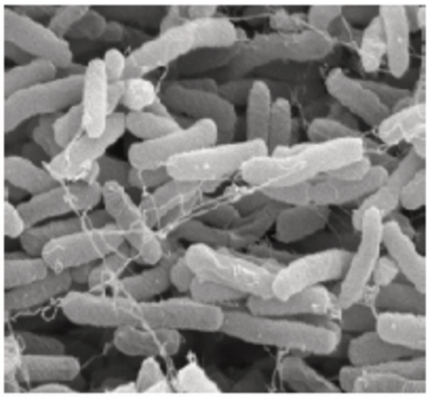
Metagenomics: a challenge for the Moore's Law

Nicolas Pons
INRA
Institut Micalis
Plateforme MetaQuant
Jouy-en-Josas, France

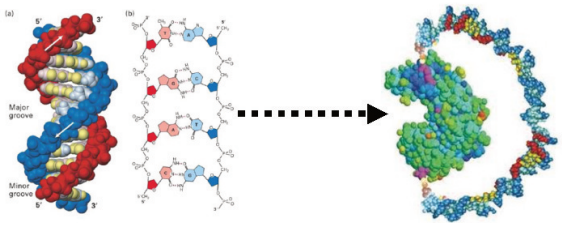
INRA

9 mai 2011

Plateforme Meta Quant

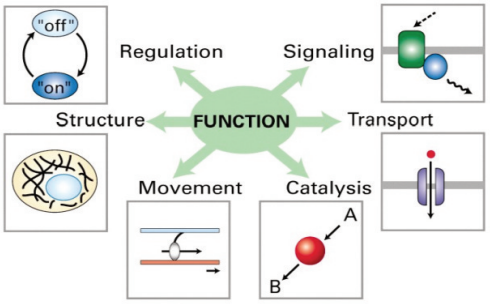


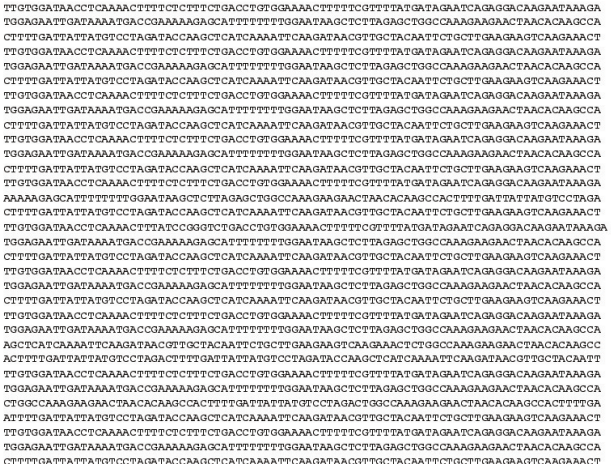
Genetic information



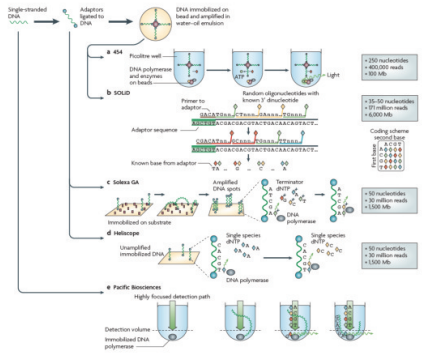
DNA Protein

Functions in the cell



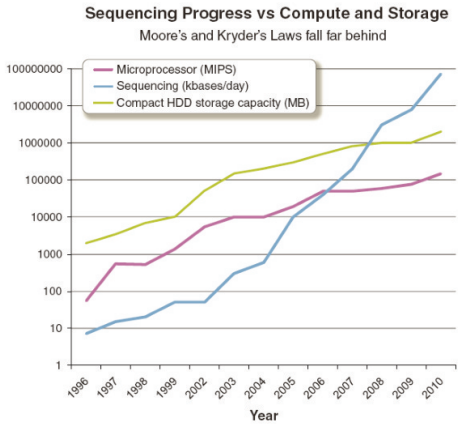


Next Generation Sequencing



Different sequencing instruments

Instrument	Run time[a]	Millions of reads/run	Bases/read[b]	Yield Mb/run	Reagent cost/run[c]	Reagent cost/Mb	Minimum unit cost (% run)[d]
3730xl (capillary)	2 h	0.000096	650	0.06	$96	$1500	$6 (1%)
Ion Torrent – '314'chip	2 h	0.10	100	>10	$500	<$50	~$750 (100%)
454 GS Jr. Titanium	10 h	0.10	400	50	$1100	$22	$1500 (100%)
Starlight*	†	~0.01	>1000	†	†	†	†
PacBio RS	0.5–2 h	0.01	860–1100	5–10	$110–900	$11–180	†
454 FLX Titanium	10 h	1	400	500	$6200	$12.4	$2000 (10%)
454 FLX+*	18–20 h	1	700	900	$6200	$7	$2000 (10%)
Ion Torrent – '316'chip*	2 h	1	>100	>100	$750	<$7.5	~$1000 (100%)
Helicos[f]	N/A	800	35	28 000	N/A	NA	$1100 (2%)
Ion Torrent – '318'chip*	2 h	4–8	>100	>1000	~$925	~$0.93	~$1200 (100%)
Illumina MiSeq*	26 h	3.4	150 + 150	1020	$750	$0.74	~$1000 (100%)
Illumina iScanSQ	8 days	250	100 + 100	50 000	$10 220	$0.20	$3000 (14%)
Illumina GAIIx	14 days	320	150 + 150	96 000	$11 524	$0.12	$3200 (14%)
SOLiD – 4	12 days	>840[g]	50 + 35	71 400	$8128	<$0.11	$2500 (12%)
Illumina HiSeq 1000	8 days	500	100 + 100	100 000	$10 220	$0.10	$3000 (12%)
Illumina HiSeq 2000	8 days	1000	100 + 100	200 000	$20 120[h]	$0.10	$3000 (6%)
SOLiD – 5500 (PI)*	8 days	>700[g]	75 + 35	77 000	$6101	<$0.08	$2000 (12%)
SOLiD – 5500xl (dhq)*	8 days	>1410[g]	75 + 35	155 100	$10 503[h]	<$0.07	$2000 (12%)
Illumina HiSeq 2000 – v3[i*]	10 days	≤3000	100 + 100	≤600 000	$23 470[h]	≥$0.04	~$3500 (6%)



From kilobytes to terabytes !!!

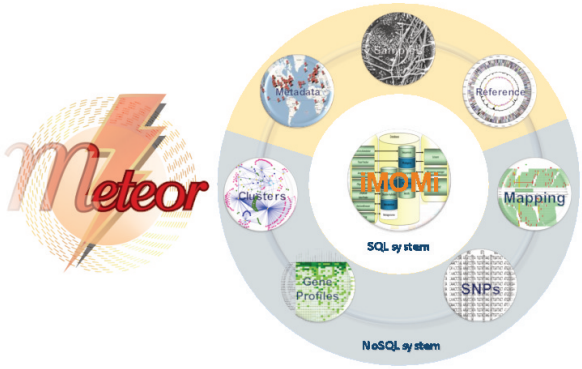
Year	Sequenced object	Sequencing unit
1990	Genes & operons	Kbase
1995	Bacterial genome	Mbase
2001	Human genome	Gbase
2010	Human metagenome	Tbase

Amount of sequence we generate has increased 10^9 times in 20 years, greatly exceeding the Moore's law

New bioinformatics challenges

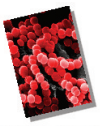
- Data managment
- Data analysis
- Data browsing

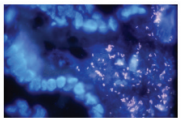
METEOR around iMOMi framework



The human intestinal microbiota is a forgotten organ…

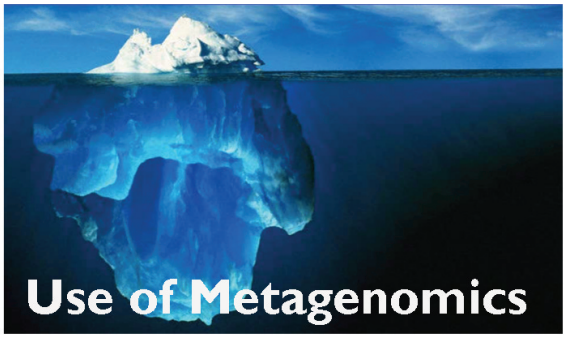
- ✓ 100 trillion microorganisms ; 10-fold more cells than the human body; 2 kg of mass!
- ✓ Interface between food and epithelium
- ✓ In contact with the 1st pool of immune cells and the 2nd pool of neural cells of the body





…with a major role in health & disease !

Most of microorganisms are unknown and uncultivable…



What is metagenomics ?

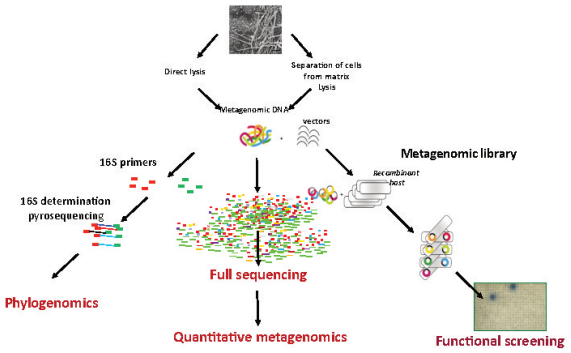
Metagenome

can be defined as the ensemble of genes of the microbes from a given ecological niche.

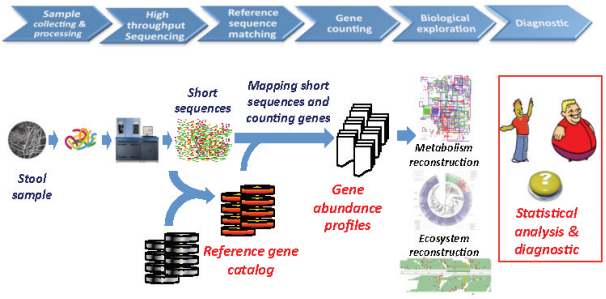
Metagenomics

allows to characterize composition, properties and dynamics of a microbiome by studying the metagenome.

Different types of metagenomics



Quantitative metagenomics pipeline



A powerful microscope!

The MetaHIT project (Dr. S.D. Ehrlich)

MetaHIT

✓ Establishing a reference gene catalog by metagenomic & genomic sequencing of the Human GI tract microbes

✓ Developing generic tools for profiling the GI tract microbiota genes : arrays and high throughput DNA sequencing

Using Quantitative Metagenomics to search associations of microbial genes and chronic disease in Obesity and Inflammatory Bowel Diseases

Illumina sequencing

BGI

Samples	124 individuals (85 Danes, 39 Spaniards)	
Library type	15 samples	200bp
	109 samples	140bp
		350bp
Sequencing type	Paired-end (PE) sequencing	
Read length (bp)	45 b (15 samples) 75 b (109 samples)	
Tags per sample	31million ±0.5 million	

In total, ~0.58 Terabase sequence

BIG DATA

Wang Jun et al.

Our other genome : the human intestinal metagenome

March 2010

3.3 million bacterial gene catalog : 150-fold human genome
Each individual has ~ 540000 of the 3.3 million genes

85% of abundant gut genes from a cohort of 124 individuals

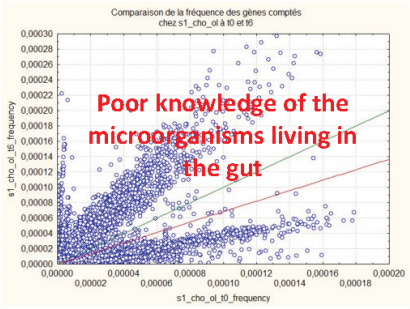
70-86% of genes from the US & Japanese studies

A complementary study : Micro-Obes (Dr. J. Doré)

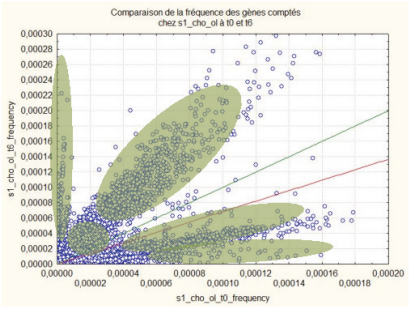
Follow the microbiota dynamic and identify signatures of obesity and nutritional transition in the intestinal microbiote

- DNA sequencing (SOLiD) of **215** faecal samples of **49** obese individuals sequenced at **3** differents time-points (**t0, w6, w12** and extra-time)
  → 300 Gbases sequenced
- Read projection against the MetaHIT gene catalog (Qin et al., 2010) (3.3 millions genes)
  →150 Gbases projected

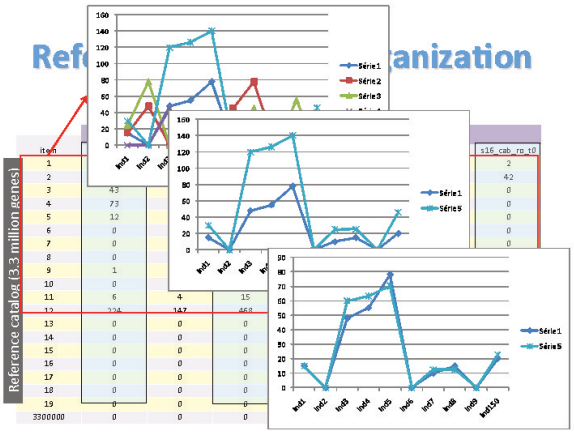
Dynamic of the intestinal microbiota



Dynamic of the intestinal microbiota



Ref anization



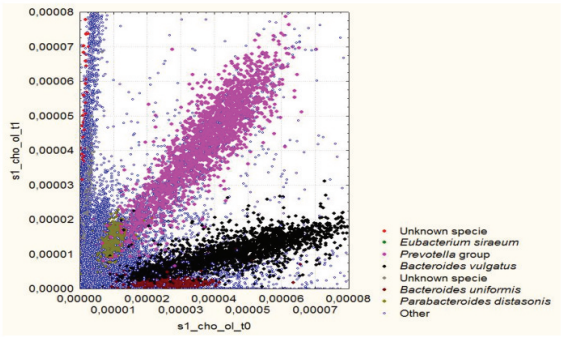
Reference catalog reorganization

- Hierarchical descendant graph & DAPC clustering (Almeida et al., 2011)
  - By computation of spearman correlation for each couple of gene profile
  - $3.3^{E}6 \times 363 \rightarrow 5^{E}12$ correlations to calculate
  - With one CPU : **more than a year to do it…**
- MetaProf (Boumezbeur, Arslan et al., 2011)
  - CUDA programming
  - 1H10 in 40 GPU

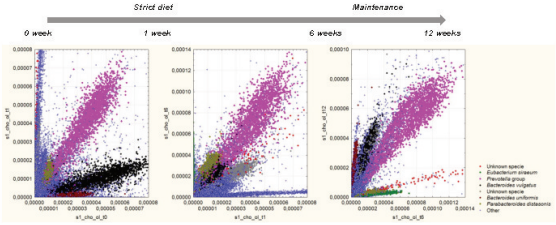
Reference catalog reorganization

- **17317 clusters** from 10 to 59353 genes (**~2.34M genes**, ~71%)
- 628 meta-species with more than 1000 genes

Dynamic of the intestinal microbiota



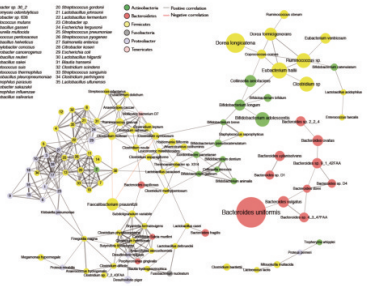
Dynamic of the intestinal microbiota



Important variation of gene composition during the first part of the diet
More stable composition in the second part (stabilization)

Co-variation of bacterial species

Network of relations between the species (circles represent species, lines the co-variance, with R≥ 0.4)



Particular constellations of bacterial species in individuals

DTU, Brunak et al.

ARTICLE

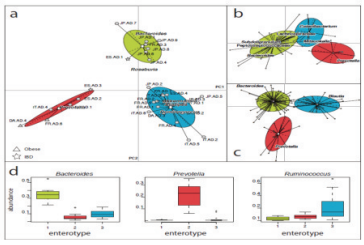
doi:10.1038/nature09944

Enterotypes of the human gut microbiome

Manimozhiyan Arumugam[1*], Jeroen Raes[2,3*], Eric Pelletier[3,4,5], Denis Le Paslier[3,4,5], Takuji Yamada[1], Daniel R. Mende[1], Gabriel R. Fernandes[2,6], Julien Tap[2,7], Thomas Bruls[3,4,5], Jean-Michel Batto[7], Marcelo Bertalan[8], Natalia Borruel[9], Francesc Casellas[9], Leyden Fernandez[10], Laurent Gautier[8], Torben Hansen[11,12], Masahira Hattori[13], Tetsuya Hayashi[14], Michiel Kleerebezem[15], Ken Kurokawa[16], Marion Leclerc[2], Florence Levenez[2], Chaysavanh Manichanh[9], H. Bjorn Nielsen[8], Trine Nielsen[11], Nicolas Pons[7], Julie Poulain[3], Junjie Qin[17], Thomas Sicheritz-Ponten[18], Sebastian Tims[15], David Torrents[19,20], Edgardo Ugarte[9], Erwin G. Zoetendal[15], Jun Wang[17,21], Francisco Guarner[9], Oluf Pedersen[11,12,22], Willem M. de Vos[15,23], Soren Brunak[8], Joel Dore[2], MetaHIT Consortium†, Jean Weissenbach[3,4,5], S. Dusko Ehrlich[3] & Peer Bork[1,25]

May 2011

Europeans, Americans, Asians. n=33; Sanger



Danes n=85; Illumina

US n=154; 454

Enterotypes can be likened to blood groups but the reasons for their existence remains to be elucidated

They should allow patient stratification & aid to develop personalized medicine and nutrition

Where do these studies lead to and when?

- **Diagnostic & prognostic tests – soon**
  - arrays, sequencing, Q-PCR; immunomarkers (?)
- **Better treatments – next**
  - personalized medicine
- **Novel treatments – last**
  - modulation of microbiota
    - Promoters
    - Inhibitors
  - transplantation of microbiota

Dr. S.D. Ehrlich :

"Take-home message:
- Our other genome has much more variability than the first one
- Personalized medicine should target it"

Conclusion

- **Data deluge is not finished**
  - 3rd generation sequencing
  - In our lab: 2 SOLiD 5500xl sequencers the next week
- **Data managing**
  - dCache / HDF5
  - Cloud
  - International repository
- **Data analysis**
  - GPU oriented programming (OpenGPU project)
- **Data browsing**

2.8 Patrick Furhmann (DESY-Hamburg)

dCache : scaling out affordable storage.

The presentation will briefly walk through the various facets of the dCache storage technology and its supporting collaborations. Functional objectives will be discussed, as well as some bits and pieces of the technical implementation. The presenter will touch upon the results of an ongoing detailed evaluation of supported file access protocols at the DESY Grid-Lab facility including a discussing on application level behaviour and pitfalls. Finally the most prominent dCache deployments will be presented and the involvement of dCache in other projects.

DCACHE.ORG DCACHE.ORG DCACHE.ORG

dCache, large scale-out affordable storage

Patrick Fuhrmann
On behave of the team

DCACHE.ORG DCACHE.ORG DCACHE.ORG

Content

- People and Funding
- Deployment
- Supported protocols
  - Data access
  - Storage control
- Managed storage
- Design facts

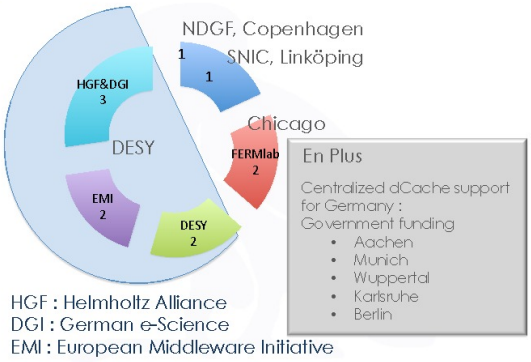
DCACHE.ORG DCACHE.ORG DCACHE.ORG

People

Funding

DCACHE.ORG DCACHE.ORG DCACHE.ORG

People and Funding



HGF : Helmholtz Alliance
DGI : German e-Science
EMI : European Middleware Initiative

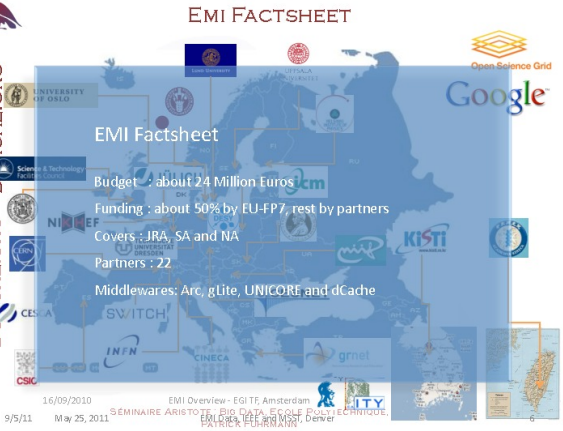
DCACHE.ORG DCACHE.ORG DCACHE.ORG

Two words on EMI

European Middleware Initiative

DCACHE.ORG DCACHE.ORG DCACHE.ORG

EMI Factsheet

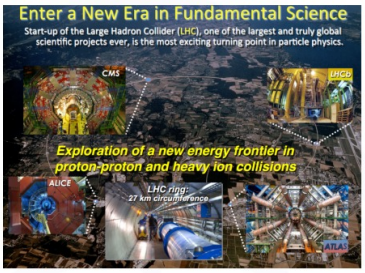


Deployment

- WLCG
- Others

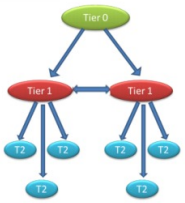
WLCG

WLCG : World Wide LHC computing Grid
LHC : Large Hadron Collider



WLCG : Specification

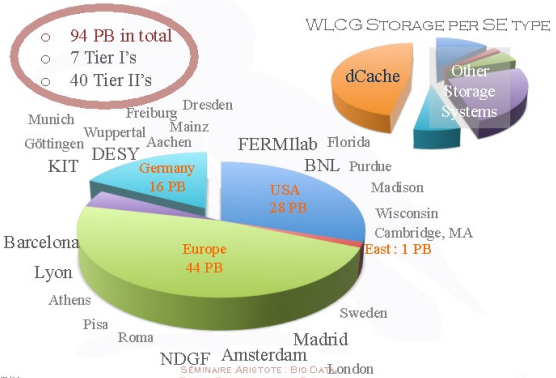
- Trying to find out why we are heavy (Higgs)
- Maybe more about dark matter and dark energy
- Producing about 15 PBytes per year
- "Raw data" is stored at CERN and at least at another Site (Tier I)



Legend
Tier 0 : CERN
Tier 1 : Counties (11)
Tier 2 : about 200

Status : dCache deployment

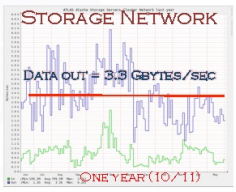
- 94 PB in total
- 7 Tier I's
- 40 Tier II's



Largest dCache Installation (BNL)
Brookhaven National Lab (New York)

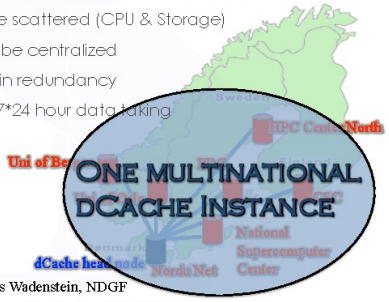
Information provided by Hironori Ito (BNL)

- 80 Million files in total
- 85 storage hosts with about 600 pools. (dCache storage unit)
- Total space on disk : 8.8 PBytes (Used are 7 PBytes)
- Total space on tape (HPSS, IBM) : 2.5 PBytes



Most interesting deployment

- The 7 biggest Nordic Computer centers form a single Tier I
- Many different tape back systems in different countries.
- Resources are scattered (CPU & Storage)
- Services can be centralized
- Advantages in redundancy
- Especially in 7*24 hour data taking



Slide stolen from Mattias Wadenstein, NDGF

Other Deployments

Historically

DESY :
HERA (Zeus, H1)

FERMILAB :
Tevatron (CDF)
Sloan Digital Sky Survey

DCACHE.ORG DCACHE.ORG

New data intensive communities

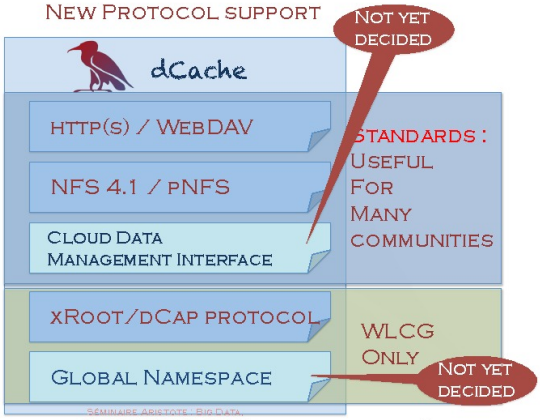
Low Frequency Array
Is using dCache at SARA (Amsterdam) and Jülich (Germany)

Center for Free Electron Laser Science
Would like to use dCache at the DESY storage.

Swedish National Infrastructure for Computing
Using dCache for Swedish academic purposes.

Supported Protocols

New Protocol support



Protocol support : WebDAV

- Very useful for new (non-LHC) communities.
- ITEF Standard
- Allows "File system like" access with
  - Mac OS
  - Linux
  - Windows

NFS v 4.1 / pNFS

My favored Topic

Protocol support : NFSv4.1 / pNFS

center for information technology integration

CITI, at the University of Michigan, is funded by major storage providers to coordinate the pNFS effort and provide reference implementations.

Industry Support - Implementations

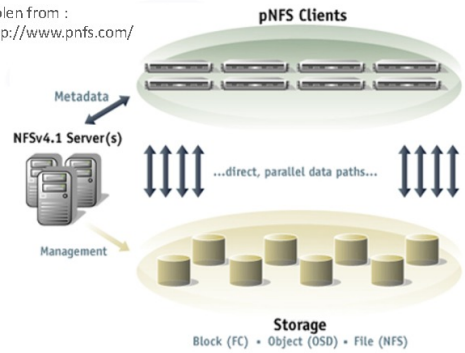
- Clients
  - Linux
  - Sun (Solaris)
- Servers
  - Desy
  - EMC
  - IBM
  - Linux
  - NetApp
  - Panasas
  - Sun (Solaris)

Several other implementations have been tested at Bake-a-thons and Connectathons

Group meets three times a year to check interoperability.

Protocol support : NFSv4.1 / pNFS

Stolen from : http://www.pnfs.com/



Protocol support : NFSv4.1 / pNFS

Stolen from : http://www.pnfs.com/

Benefits of Parallel I/O

- ✓ Delivers Very High Application Performance
- ✓ Allows for Massive Scalability without diminished performance

Benefits of NFS (or most any standard)

- Ensures Interoperability among vendor solutions
- Allows Choice of best-of-breed products
- Eliminates Risks of deploying proprietary technology

Protocol support : NFSv4.1 / pNFS

Simplicity

- ✓ Regular mount-point and real POSIX I/O
- ✓ Can be used by unmodified applications (e.g. Mathematica..)
- ✓ Data client provided by the OS vendor
- ✓ Smart caching (block caching) development done by OS vendors

Performance

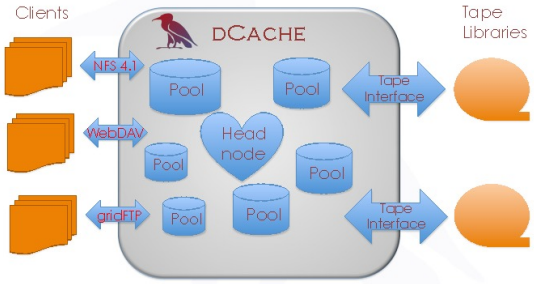
- ✓ pNFS : parallel NFS (first version of NFS which support multiple data servers)
- ✓ Clever protocols , e.g. Compound Requests

dCache is

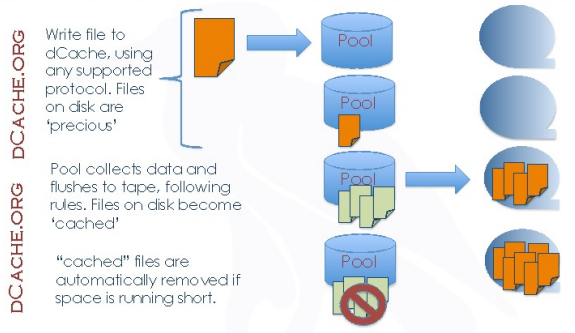
Managed Storage

- Manual Storage Management
  - The Storage Resource Manager (SRM)
  - Storage Migration Module
- Automatic Storage Management
  - Storage Attribute by directory
  - Hot Spot detection
  - Resilient Manager

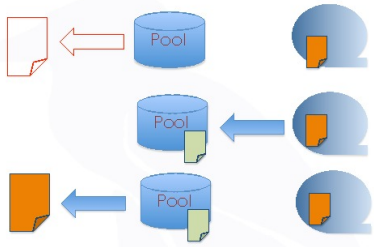
dCache Idea



Normal dCache file cycle (Storing data)



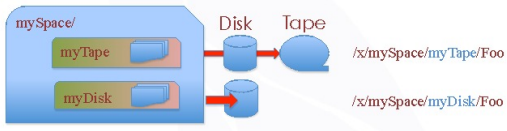
Normal dCache file cycle (Retrieving data)



In case of a '*cache miss*': The file is automatically restored from tape and subsequently delivered to clients, using the selected protocol.

Managed Storage

Final destination determined by directory



Managed Storage : SRM

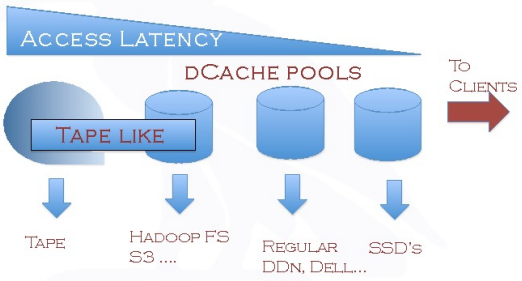
The storage resource manager protocol

- SRM 2.2 : defined by the Open Grid Forum (OGF)
  - ✓ Defines storage media (Disk/Tape)
  - ✓ Can use "Spaces" (similar to AMAZON buckets) with attributes (disk, tape, size)
  - ✓ Pin / Unpin files
  - ✓ Bring Online file(s), in preparation for a read.
  - ✓ Remote secure protocol with many implementations

More Managed Storage

- Automatic file replication on 'hot spot' detection
  - ✓ If a pool is used heavily, dCache starts to spread files from this pool to other (lazy) pools.
- Resilient manager
  - ✓ On basis of a pool set, a minimum and maximum number of replicas for all files can be defined.
  - ✓ dCache automatically adjusts the replicas if pools go down or are scheduled for maintenance.
- Migration Module
  - ✓ Files and be shuffled around between pools (by rules) to allow to spread load or decommission pools.

dCache allows to make use of different storage capabilities

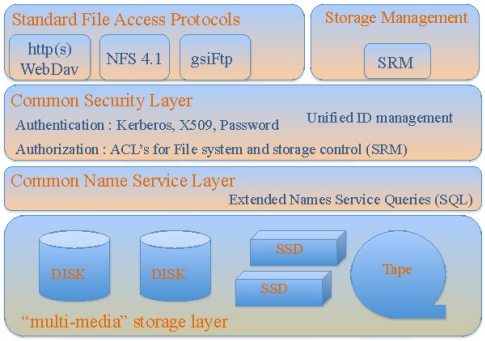


DCACHE.ORG DCACHE.ORG DCACHE.ORG

Design Details

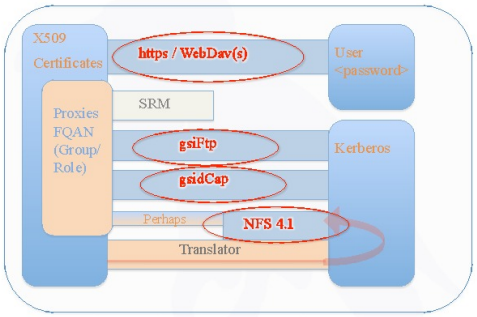
DCACHE.ORG DCACHE.ORG DCACHE.ORG

dCache, in layers



DCACHE.ORG DCACHE.ORG DCACHE.ORG

Authentication



DCACHE.ORG DCACHE.ORG DCACHE.ORG

No Secrets Anymore

- All Java
- Name space abstraction
  - Legacy implementation (PNFS) or
  - New Implementation (any JDBC DB, def. postgres)
- Component communication via message passing:
  - Private Protocol (Cells) or
  - Java Messaging Service (JMS)
- Scalable components : Protocol Endpoints and data pools
- Single Point of failures : namespace and pool/space manager
- With 1.9.12 (2nd Golden Release) : Very nice configuration system

DCACHE.ORG DCACHE.ORG DCACHE.ORG

Conclusion

- dCache is about storing, accessing and managing huge amounts of data.
- Depending on the configuration (resilient manager) you may use cheap hardware.
- Historically tuned for HEP and WLCG
- For about 2 year focusing on more communities, which have been committing themselves to standards (web 2.0)
- dCache collaboration nicely distributed amongst Europe and the US.
- Funding spread amongst different bodies. (e.g EMI)
- More contributions/contributors welcome.

DCACHE.ORG DCACHE.ORG DCACHE.ORG

Further Reading

www.dCache.org

2.9 Marie-Luce Picard (EDF R&D et ENST-Bilab)

Données massives pour les *smart-grids*

De nombreux projets *smart-grids* voient le jour à travers le monde, motivés par des aspects régulatoires, des contraintes économiques ou la prise en compte de besoins environnementaux ou sociaux. Ces projets reposent sur le déploiement de compteurs communicants et la mise en place d'une infrastructure de communication adéquate. Mais il ne s'agit là que de la première étape de la mutation technique et économique du secteur énergétique. Cette vision long terme de la problématique des réseaux intelligents sous-tend une capacité à gérer et traiter de larges volumes de données, provenant en particulier des compteurs intelligents ou encore de différents capteurs sur le réseau. Dans cette perspective, un certain nombre de travaux ont été menés à EDF R&D et seront présentés dans cet exposé, en particulier : le stockage de grandes quantités de séries temporelles, le traitement temps-réel de courbes de charge imparfaites, les perspectives d'évolution des approches de prévision de consommation en présence de données individuelles massives.

Données massives pour les smart-grids

Marie-Luce PICARD
EDF R&D
marie-luce.picard@edf.fr

9 Juin 2011

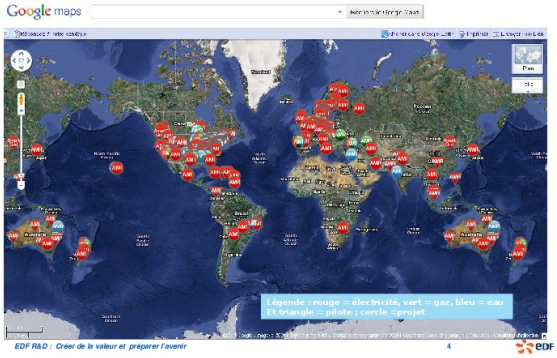
EDF
CHANGER L'ÉNERGIE ENSEMBLE

Sommaire

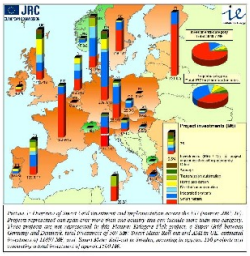
1. Smart-Grids : systèmes électriques intelligents
2. Données et Smart-Grids
3. Travaux menés à EDF R&D
4. Conclusion

Smart Grids : systèmes électriques intelligents

Les réflexions autour des Smart-Grids voient le jour partout dans le monde ...



Un enjeu majeur ... De nombreux investissements



- Les incitations des pouvoirs publics encouragent l'innovation et les expérimentations en la matière (CRE, Ademe, UE ...)
- Des investissements importants ont été réalisés par des acteurs TIC ou industriels
- Des initiatives de mise en œuvre ont été effectuées par de nombreuses utilities

Et en particulier en France

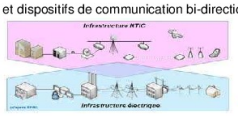
Commission de Régulation de l'Energie (CRE) : www.smartgrids-cre.fr
ERDF : linky.erdfdistribution.fr

Décret d'Août 2010, qui indique que :
- A partir du 01/01/2012 tous les nouveaux compteurs doivent être communicants
- Après le 31/12/2014 , 50% des compteurs installés doivent être communicants
- Après le 31/12/2016 , 95 % des compteurs installés doivent être communicants.

Smart-Grids : systèmes électriques intelligents

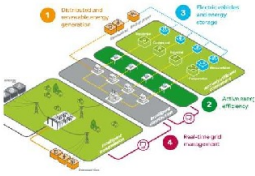
Source – Wikipedia
A **smart grid** delivers electricity from suppliers to consumers using digital technology with two-way communications to control appliances at consumers' homes to save energy, reduce cost and increase reliability and transparency. It overlays the electrical grid with an information and net metering system, and includes smart meters. Such a modernized eletricity network is being promoted by many governments as a way of addressing energy independence, global warming and emergency resilience issues.

- *«Un Système électrique intelligent est un système électrique capable **d'intégrer de manière intelligente les actions des différents utilisateurs, consommateurs et/ou producteurs** afin de maintenir une fourniture d'électricité efficace, durable, économique et sécurisée».*
  - Passage d'un modèle traditionnel centralisé à un modèle distribué, interconnecté, avec de fortes interactions (avec le client)
  - Le réseau se dote de nombreux capteurs et dispositifs de communication bi-directionnel
  - Réseau électrique + réseau informatique



Smart-Grids : systèmes électriques intelligents

- Les enjeux des Smarts-Grids :
  - Permettre notamment une gestion plus interactive de la demande et de la production décentralisée
  - Favoriser les comportements de MDE (Maîtrise de l'Énergie) : information du client, effacements, demand response …
  - Gérer les énergies intermittentes et la qualité de fourniture
  - Gérer le développement de nouveaux usages tels que le véhicule électrique

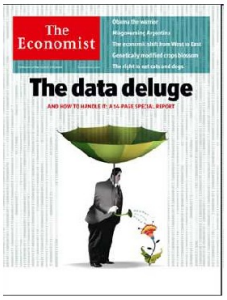


⇨ Impact sur les relations entre les acteurs
⇨ Impact sur les business models

⇨ Impact significatif sur les volumes de données générés, et donc à manipuler, parfois en temps réel … de façon distribuée
⇨ Impact sur les Systèmes d'Information et leur nécessaire agilité

Données et Smart-Grids

Données massives ?



Traitement de données massives : quelles technologies ?

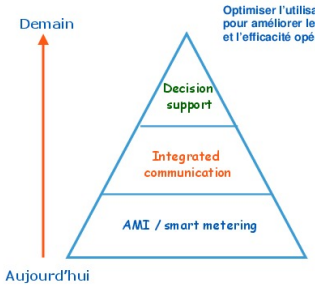
Technologies	Exemples
Très grands entrepôts de données	*Wal-mart* Entrepôts à large échelle : Teradata, Exadata, InfoSphere
Traitement de données à la volée	*CME Group* Complex Event Processing : StreamBase, IBM InfoSphere Streams
Approches distribuées pour le stockage	*Google, Facebook* Cloud computing (MapReduce) Hadoop
Data mining à très grande échelle	*Amazon* Systèmes de recommandation Data mining distribué

Qu'en est-il pour les Smart-Grids ?

- La mise en place de fonctionnalités de type Smart-Grids va entraîner une très forte augmentation du volume des données à traiter :
  - Installation des compteurs communicants (35 millions en France)
  - Installation de capteurs divers sur le réseau
  - Évolution des installations chez le clients (objets communicants …)
  - ~ ***Peta-octets*** (d'ici 5 / 10 ans)
- Ces volumes sont toutefois en-deçà de ceux rencontrés par d'autres secteurs d'activité qui ont d'ores et déjà à leur disposition différents outils technologiques
- Néanmoins on entrevoit la mise en place de traitements complexes :
  - De par la nature des données : des séries temporelles
  - De par leur caractère distribué et de la nécessité de traitements à différentes échelles
  - De par les contraintes temps réel pour certains besoins

A Grid is a Grid: What Does Scale Have to Do with It?(SDG&E)

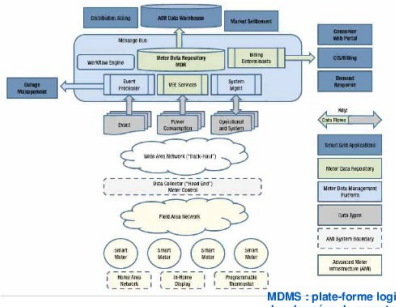
Fonctionnalités Smart-Grids : roadmap



Architectures pour les Smart-Grids

- L'expérimentation à grande échelle de projets Smart-Grids est une réalité (aux Etats-Unis)
  - **Projets à forte composante informatique**
  - Projets sociétaux (acceptabilité, privacy)
  - Exemples : PG&E (5M), SDG&E (1,5 M)
- Architectures mises en place :
  - Basées sur des solutions SI matures et standards (intégrations autour d'un bus d'entreprise)
  - Notion de MDMS (Metering Data Management System)
  - Fonctionnalités :
    - Canal batch pour les données horaires mises à disposition le lendemain sur portails web
    - Canal temps réel pour alertes
    - Mise à disposition des données du MDMS pour la facturation
    - Alimentation d'un entrepôt de données pour une historisation des données de comptage

Architectures pour les Smart-Grids (2)



MDMS : plate-forme logicielle qui acquiert des données de comptage depuis de nombreuses sources et les mets à disposition, après intégration, synchronisation et nettoyage, auprès de nombreuses cibles.

Architectures pour les Smart-Grids (3)

- Architectures à moyen terme :
  - Elles intégreront une intelligence distribuée (routeurs intelligents à capacité de traitement, bases de données distribuées, traitements des données et évènements temps réel (CEP) distribués) qui pourra être pilotée par un système centralisé
  - Traitement conjoint des données de comptage et des données liées au fonctionnement du réseau
    - Fonctions avancées de conduite et d'exploitation du réseau (observabilité, réseaux auto-cicatrisants ...)
    - Gestion de la demande, prix temps réels
    - Intégration étroite des énergies renouvelables
    - Optimisations locales
  - ***Systèmes de systèmes***

The necessary functions remain the same, the key issue is manage the complexity to support the necessary business capabilities at any scale as well as manage the separation of responsabilities to avoid "dueling" control systems (SDG&E)

Travaux menés à EDF R&D

Données, infrastructures, technologies de traitement

- Aujourd'hui :
  - Des panels de courbes de charges et des profils
  - Délai important de récupération et de mise à disposition des données
- Demain :
  - Des données de comptage à fine granularité, individuelles ou locales
  - Faible latence entre les systèmes opérationnels et les systèmes d'information décisionnels
- Des technologies de traitement de données « massives » permettant d'envisager une exploitation à grande échelle des données issues des infrastructures communicantes :
  - Gros entrepôts de données, décisionnel temps réel
  - Traitements de données à la volée (« Complex Event Processing », flux de données)
  - Traitements et décisions locales

Stockage Massif de Courbes de charge

- Objectif : montrer la faisabilité d'un stockage massif de courbes de charges rendues disponibles pour un certain nombre de traitements (plus ou moins complexes, plus ou moins concurrents, avec une latence variable selon les besoins)
  - Données : courbes individuelles, données météo, informations contractuelles, données réseau
    - 1 relevé pas 10 mn / 35 millions de clients / an
    - Volumes de données annuels
      - 1800 milliards de relevés ; 600 To de données non compressées
    - Volumes de données journalier
      - 5 milliards de relevés ; ~2 To

  - Grandes fonctions évaluées
    - Alimentation des données (par batch & par flux)
    - Rapprochement avec les données CRM
    - Pré-traitement des données
      - Synchronisation temporelle
      - Détection et correction d'anomalies
      - Changement de modes de représentation
    - Traitements
      - Calcul de synchrones par sous-population
      - Simulation de facturation (agile)
      - Simulation de traitements simultanés
        - Sélection d'une courbe sur motifs
        - Requêtes analytiques, BI
  - Critères d'évaluation
    - Hautes performances
      - Alimentation des données, accès aux données, répartition des charges
    - Haute disponibilité
      - Tolérance aux pannes, redondance et secours
    - Extensibilité
      - Fonctionnelle et horizontale
    - Type d'architecture : share-nothing ou everything

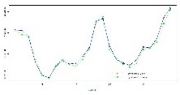
Stockage Massif de Courbes de charge

- Approches envisagées :
  - Approches distribuées :
    - POC interne, basé sur l'éco-système Hadoop en cours
    - Stockage et traitement distribué de type MapReduce
    - La maturité industrielle de ces approches n'est pas encore complètement atteinte
  - Approches relationnelles classiques :
    - Travaux réalisés avec des partenaires
    - Appliances - POCs en cours
  - Data Historian

- Réalisation d'un générateur de données ***CourboGen***
  - Permet la génération de courbes de charge et de données associées
  - Outil paramétrable : pas de temps, durée, qualité des données, bruit sur les courbes
  - Architecture distribuée
  - Performances : 60000 relèves / seconde / module
  - Sortie sous forme de flux de données

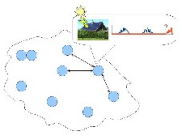
Constitution de courbes synchrones

- Objectif : exploitation / valorisation des données de comptage massives pour différents besoins réseaux, avec une faible latence
  - Constitution de courbes synchrones
    - Par segment logique (portefeuille client)
    - Par segment topologique (maille réseau)
- Points à résoudre
  - Effectuer le calcul en temps réel, ou avec une grande réactivité
  - Volumes et bande passante
  - Qualité des données (pertes d'informations, retards, dysfonctionnements)
- Éléments de réponse :
  - Architectures multi-chemins et mécanismes robustes à la duplication (« sketches »)
    - Architectures possibles et distribution de l'intelligence ?
    - Latence acceptée pour les traitements ?
  - Inscrire ce calcul dans des outils de traitement de flux de données (traiter avant de stocker, à la volée)
    - Maquette utilisant le CEP StreamBase : traitement à la volée de 300000 courbes dupliquées 6 fois ; simulation de 2 RE (95% et 5% du volume des consommations)



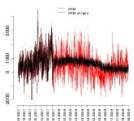
Prévision de production locale

- Objectif : pilotage de l'équilibre du réseau au niveau local (intégration des ENR, gestion des VE), optimisations locales de fonctions réseau
- Un élément de réponse : le data-mining distribué
  - Approches basées sur des agents indépendants les uns des autres qui peuvent effectuer des tâches différentes et collaborer ensemble
    - Données très distribuées, vision globale sans centraliser l'ensemble des données
    - Besoins de fouille locale, avec un contexte global (collaboration entre les tâches locales)
- Premières expérimentations :
  - Prévision de production PV locale court terme



Prévision de consommation

- Objectif : prise en compte dans les méthodes et outils de prévision de l'évolution « smart grid » (données disponibles de façon détaillée : quelle opportunité ? Besoins de prévisions à mailles plus locales, et avec une plus grande variabilité des périmètres concernés)
- Aujourd'hui :
  - Les modèles de prévision sont robustes et de (très) bonne qualité
  - Ils sont néanmoins conçus pour travailler sur un périmètre constant et un signal de consommation régulier
- Demain :
  - Périmètres variables, signaux non stationnaires (mouvements de clients, périodes difficiles à prévoir, incertitudes, intégration de la production décentralisée, nouveaux usages)
  - Quels bénéfices peut-on tirer de l'utilisation de données individuelles massives ?
    - Données arrivant sous forme de flux, à prendre en compte pour les modèles court terme (de la veille pour le lendemain) et infra-journaliers
      - Adaptativité
      - Mouvements de clients captés avec faible latence
    - Données individuelles :
      - Agrégation (clustering, échantillonnage)



Résidus de prévision sur un an
(fenêtre de taille 100, modélisation du niveau moyen et de l'effet retard en ligne)
RMSE GAM : 900 MW. RMSE GAM en ligne : 714 MW

Conclusion

Données massives pour les smart grids : impacts et enjeux

- Les projets « Smart-Grids » ont une forte composante informatique
- Ils ont et auront un impact fort sur l'évolution du secteur énergétique :
  - Un réseau informatique, et pas seulement électrique
  - NTIC et utilities
  - Une augmentation forte du volume des données à traiter sur lesquelles on va vouloir appliquer des traitements parfois complexes
  - Place majeure du / des SI (volumétrie, complexité, architecture et urbanisme, agilité)
- Passage à l'échelle et verrous :
  - Stockage de larges volumes de séries temporelles
  - Approches centralisées et/ou distribuées
  - Traitement de données à la volée et en temps réel (CEP éventuellement distribués)
  - Data-mining à large échelle :
    - Distribution et parallélisation des algorithmes de fouille (cloud)
    - Le data-mining distribué peut amener des réponses en terme de volumétrie et de respect de la privacy
    - Apprentissage en ligne

Travaux réalisés avec :

Charles Bernard
Alexis Bondu
Xavier Brossat
Yousra Chabchoub
Leeley Daio Pires Dos Santos
Alzennyr Gomes Da Silva
Veronica Gomez
Yannig Goude
Benoît Grossin
Georges Hébrail
Bruno Jacquin
Sylvie Mallet
Amandine Pierrot
David Worms

Rererences

- **MDMS** : rapport du GTM Research
  www.gtmresearch.com/report/the-emergence-of-meter-data-management-mdm
- **Smarter Energy @ IBM Research**, Brian Gaucher, Manager Smarter Energy, IBM T.J. Watson Research Center
- **Analytics and transactive control design for the Pacific Northwest Smart Grid Demonstration Project**, P. Huang, J. Kalagnanam, R. Natarajan (IBM Research Watson), D. Hammerstrom and R. Melton, Battle Memorial Institue, Pacific Northwest Division, Richland. (http://www.ieee-smartgridcomm.org/techprogram.html)
- hadoop.apache.org
- **Agrégation robuste de données en temps réel : application aux compteurs électriques communicants**, Y. Chabchoub, B. Grossin, Prix du meilleur article applicatif EGC 2011
- **Short term electricty load forecasting with adaptive GAM models**, Y. Goude, A. Pierrot, ISF 2010
- **A range of methods for electrical consumption forecasting,** X. Brossat, ISF 2010.
- « **SMAC – Stockage Massif de Courbes** », séminaire BILab du 12 Mai 2011, B. Jacquin, L. Daio Pires Dos Santos, A. Gomes Da Silva, D. Worms
- http://www.csee.umbc.edu/~hillol/Kargupta/pubs.html

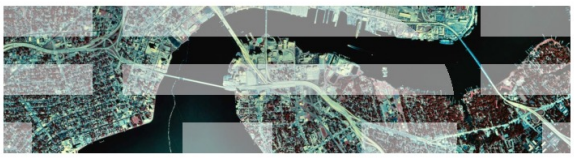
2.10 David KONOPNICKI (IBM Haïfa-Research)

Massive-scale Analytics for a Smarter Planet

Everyday, we create 2.5 quintillion bytes of data–so much that 90% of the data in the world today has been created in the last two years alone. This data comes from everywhere : from sensors used to gather climate information, posts to social media sites, digital pictures and videos posted online, transaction records of online purchases, and from cell phone GPS signals to name a few. This data is Big Data. Big Data is more than a challenge ; it is an opportunity to find insight in new and emerging types of data and to answer questions that, in the past, were beyond reach. Until now, there was no practical way to harvest this opportunity. Today, IBM's platform for Big Data opens the door to a world of possibilities, giving organizations a solution that is designed specifically with the needs of the enterprise in mind and provides the infrastructure of a Smarter Planet : intelligence is being infused into the systems and processes that make the world work—into things no one would recognize as computers : cars, appliances, roadways, power grids, clothes, even natural systems such as agriculture and waterways.

IBM

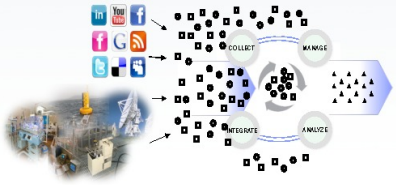
Massive Scale Analytics for a Smarter Planet



IBM

The **Big** Data Challenge

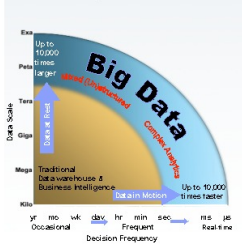
- Manage and benefit from massive and growing amounts of data
  - 44x growth in coming decade from 800,000 petabytes to 35 zettabytes
- Handle unstructured data (text/images/video), social (graph) data…
- Exploit **BIG Data** in a timely and cost effective fashion



IBM

Massive Scale Analytics

- A new set of tools is emerging around Big Data
- Internet technology for handling such Big Data is now entering the enterprise

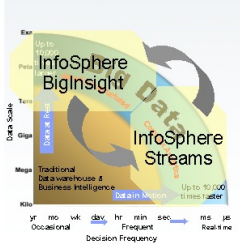


	"Traditional" Data	Big Data
Amount of data	Hundreds of Tera Bytes in large cases	Internet Scale, PetaBytes
Type of data	Mostly Relational & structured	Unstructured, Human and Sensor Generated
Deep Analytics	Statistic modeling Sampling	Predictive, Fine Grained Analytics
Affordability	Only very big Enterprises	Pay-as-you-go through cloud expansion
Insight gained	Customer segmentation	Service that knows "you"

IBM

Massive Scale Analytics

- A new set of tools is emerging around Big Data
- Internet technology for handling such Big Data is now entering the enterprise



	"Traditional" Data	Big Data
Amount of data	Hundreds of Tera Bytes in large cases	Internet Scale, PetaBytes
Type of data	Mostly Relational & structured	Unstructured, Human and Sensor Generated
Deep Analytics	Statistic modeling Sampling	Predictive, Fine Grained Analytics
Affordability	Only very big Enterprises	Pay-as-you-go through cloud expansion
Insight gained	Customer segmentation	Service that knows "you"

IBM

Massive Scale Data Analytics Problem and Solution

- How do you **scale up** applications?
  - Run jobs processing 100's of terabytes of data
  - Takes 11 days to read on 1 computer
- Need lots of cheap computers
  - Fixes speed problem (15 minutes on 1000 computers), but…
  - Reliability problems
    - In large clusters, computers fail every day
    - Cluster size is not fixed
- Need common infrastructure
  - Must be efficient and reliable

- **Infrastrucure:**
  - Google File System – *SOSP'2003*
  - **Hadoop File System**
- **Computing Paradigm:**
  - **Map-Reduce** – *OSDI'2004*
  - **Hadoop Map-Reduce**
- **Scripting Language:**
  - **Sawzall** – *Scientific Programming Journal'2005*
  - **Pig, Jaql**
- **NoSQL DB:**
  - **Big Table** – *OSDI'2006*
  - **HBase, Cassandra, Hive**
- **Workflow:**
  - **OOzie, Metatracker**

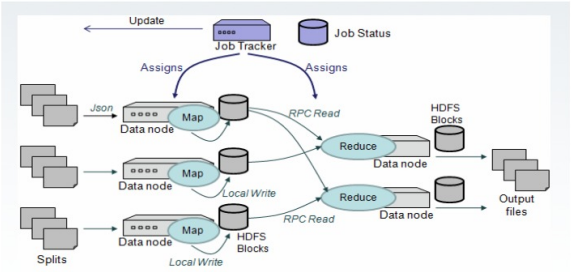
IBM

Solution

- Open Source Apache Project
- Hadoop Core includes:
  - Distributed File System - distributes data
  - Map/Reduce - distributes application
- Written in Java
- Runs on
  - Linux, Mac OS/X, Windows, and Solaris
  - Commodity hardware

IBM

Hadoop in a Nutshell



•An open-source computing platform that is both distributed and redundant, handles structured and unstructured data and supports a simple and efficient programming paradigm called map-reduce.

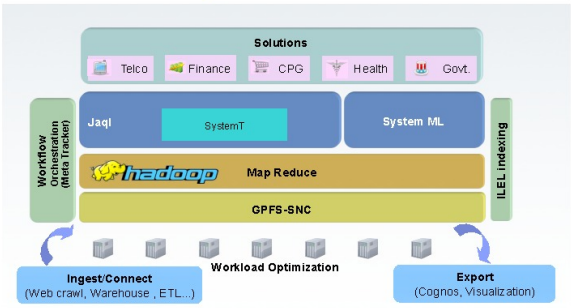
IBM

Who Uses Hadoop?

- Amazon/A9
- AOL
- Facebook
- Fox interactive media
- Google / IBM
- Netflix
- New York Times
- PowerSet (now Microsoft)
- Quantcast
- Rackspace/Mailtrust
- Veoh
- Yahoo!
- More at http://wiki.apache.org/hadoop/PoweredBy

IBM

IBM's Massive Scale Analytics Architecture



Innovation above, underneath and around Hadoop

IBM

Products Impact

- Cognos Consumer Insight: Mining social media helps marketing professionals transform customer relationships by analyzing sentiment, affinity and evolving topics in social media sites.
- Infosphere BigInsight: Bringing the power of Hadoop to the Enterprise

IBM

Smarter Planet: Something profound is happening…

INSTRUMENTED
We now have the ability to measure, sense and see the exact condition of practically everything.

INTERCONNECTED
People, systems and objects can communicate and interact with each other in entirely new ways.

INTELLIGENT
We can respond to changes quickly and accurately, and get better results by predicting and optimizing for future events.

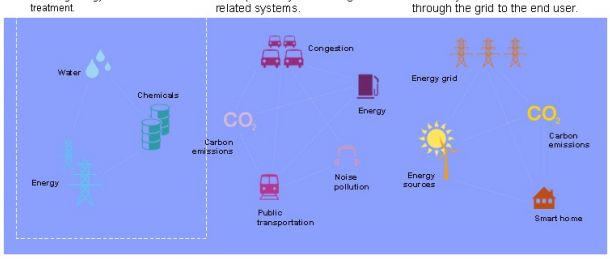
IBM

Intelligent systems that gather, synthesize and apply information will change the way entire industries operate.

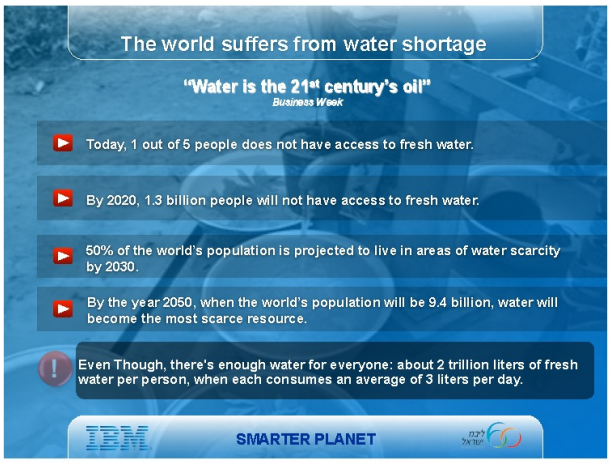
Smart water
Apply monitoring and management technologies to help optimize the availability, delivery, use, and quality of water as well as related systems including energy and chemical treatment.

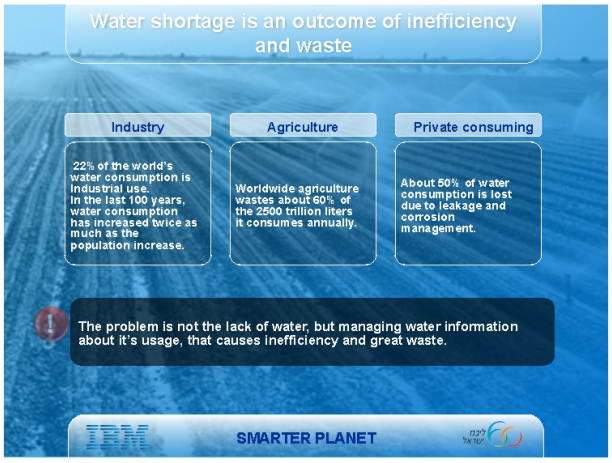
Smart traffic
Use real-time traffic prediction and dynamic tolling to reduce congestion and its byproducts while positively influencing related systems.

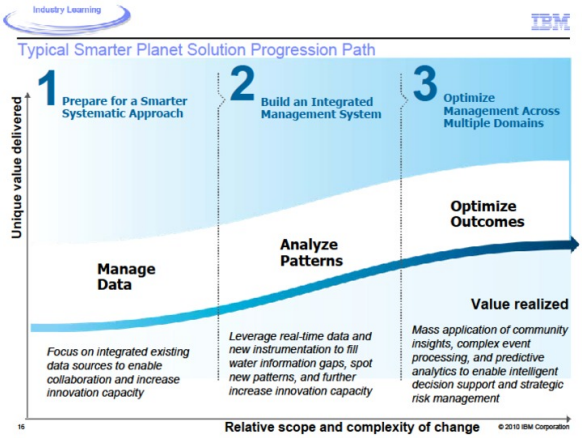
Smart energy
Analyze customer usage and provide customized products and services that help to boost efficiency from the source through the grid to the end user.




How much water do you need to..
20 liters of water to manufacture one sheet of paper
70 liters of water to create one apple
91 liters of water to manufacture one plastic can
140 liters of water to manufacture one cup of coffee
15,500 liters of water to create one kilo of beef
40 liters of water to create a loaf of bread
80 liters of water per dollar of an industrial product
120 liters of water to create a glass of wine
1300 liters of water to manufacture one kilo of wheat
10855 liter of waters to manufacture one pair of jeans
IBM
SMARTER PLANET


The world suffers from water shortage
"Water is the 21st century's oil"
Business Week
Today, 1 out of 5 people does not have access to fresh water.
By 2020, 1.3 billion people will not have access to fresh water.
50% of the world's population is projected to live in areas of water scarcity by 2030.
By the year 2050, when the world's population will be 9.4 billion, water will become the most scarce resource.
Even Though, there's enough water for everyone: about 2 trillion liters of fresh water per person, when each consumes an average of 3 liters per day.
IBM
SMARTER PLANET


Water shortage is an outcome of inefficiency and waste
Industry
22% of the world's water consumption is industrial use. In the last 100 years, water consumption has increased twice as much as the population increase.
Agriculture
Worldwide agriculture wastes about 60% of the 2500 trillion liters it consumes annually.
Private consuming
About 50% of water consumption is lost due to leakage and corrosion management.
The problem is not the lack of water, but managing water information about it's usage, that causes inefficiency and great waste.
IBM
SMARTER PLANET


Industry Learning
IBM
Typical Smarter Planet Solution Progression Path
1 Prepare for a Smarter Systematic Approach
2 Build an Integrated Management System
3 Optimize Management Across Multiple Domains
Unique value delivered
Manage Data
Analyze Patterns
Optimize Outcomes
Value realized
Focus on integrated existing data sources to enable collaboration and increase innovation capacity
Leverage real-time data and new instrumentation to fill water information gaps, spot new patterns, and further increase innovation capacity
Mass application of community insights, complex event processing, and predictive analytics to enable intelligent decision support and strategic risk management
Relative scope and complexity of change
© 2010 IBM Corporation


IBM
Addressing Non-Revenue Water using Analytics and Optimization
Leakage or Theft Detection at the Residential, public buildings.
Understand usage patterns and detect anomalies for low and high consumption to detect leakage, theft or faulty meters
Leakage Detection at the Network Level using optimization
Find "optimal" location of leak(s) to explain difference between actual measurements and model predicted measurements
Leakage Reduction using Dynamic Pressure Control
Create optimization model to adjust the pressure dynamically so that only the required flow will be supplied yielding cost reduction in energy and water achieved.
Optimal Valve Placement for Pressure Reduction
Use an optimization model to find the optimal number of valves, and their location, so as to enable the most effective pressure management
17
© 2009 IBM Corporation


Why now?
Leaders focus on credit crisis
Roubini warns of double-dip recession: rep
Consumers Postpone Purchases as Recession Deepens
One certainty: Gas will go up
Where will it end?
World Bank warns of social unrest
news.bbc.co.uk - 24 May 2009
Why The Telcos Are Doomed
"Because we must"
WHAT WENT WRONG WITH ECONOMICS?
Climate Change
When glaciers start moving
U.S. must "adapt" to changing climate
reuters
Can Digital Health Protect Your Privacy?
CNN MONEY
Enemies in the Media
IBM
SMARTER PLANET

IBM

IBM Customer Modeler

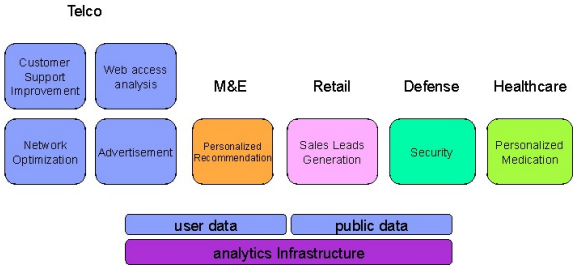
IBM

IBM Customer Modeler

- A components library to analyze customer behavioral data and enable new insights and business scenarios based on web behavioral data
- Behavioral data:
  - Web Search activities (internal and external)
  - Web Browsing (internal and external)
  - Customer generated content (blogs, twitter)
  - Products browsed, bought
  - Location, context
- Other sources:
  - Enterprise:
    - Product Catalogs
    - Web site metadata
    - Customer data, segmentations, taxonomies
  - Public
    - Wikipedia
    - Open directory projects

IBM

IBM Customer Modeler in Industries



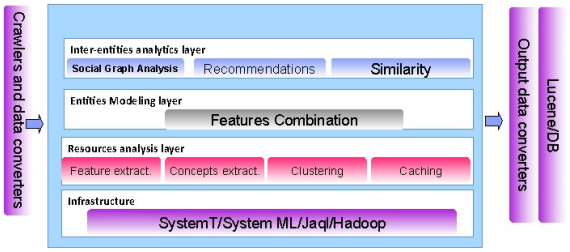
IBM

Scenarios

- Network Optimization: Learning customer browsing activities in order to predict future data access and optimize caching
- Advertisement: Learning customer interests though browsing activities in order to optimize ads targeting
- Customer support improvement: Learning customer search and browsing activities in order to discover information need and improve interaction, knowledge bases
- Web access analysis: Learning customer interests and clustering to detect trends
- Personalized Recommendation: Learning customer preferences from consumption data and providing personalized recommendations
- Sales leads generation: Analyzing customers (private and companies) generated content to discover opportunities
- Healthcare: Analyzing social media to find data relevant to particular patients

IBM

BigInsight Library for Industry Scenarios



IBM

Components in Details

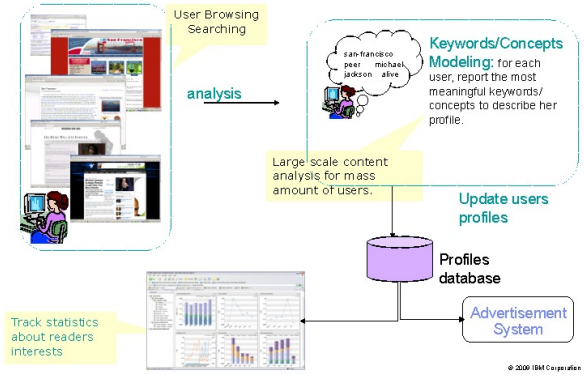
IBM

Components

- Web Browsing/Searching Analytics
- Personalized Recommendations
- Finding Influencers
- Finding Similarities

IBM

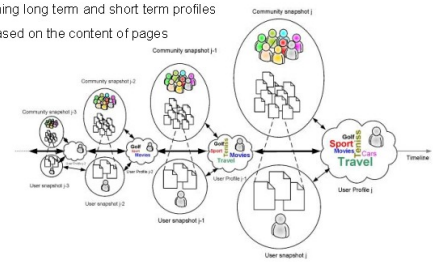
Mining Web Behavior



IBM

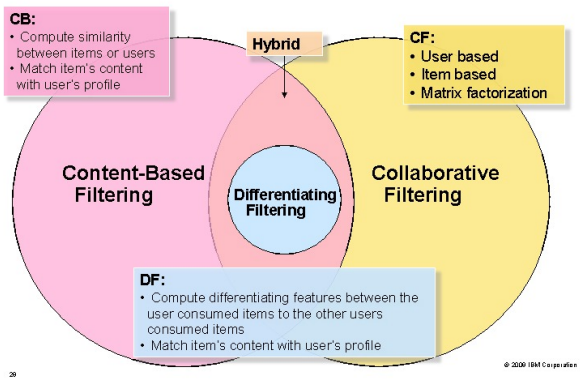
Mining Web Behavior

- Source: Internal/External Browsing/Searching History
- Basic concept: finding what statistically differentiates a particular user from the rest of the population
- Maintaining and combining long term and short term profiles
- Can be implemented based on the content of pages or based on metadata



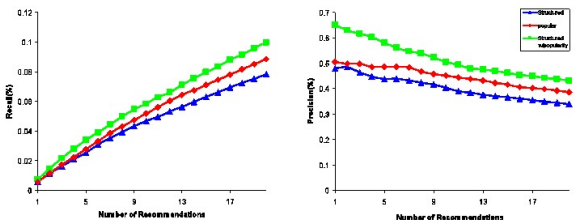
IBM

Preferences Prediction for recommendations



IBM

Enhanced Popularity

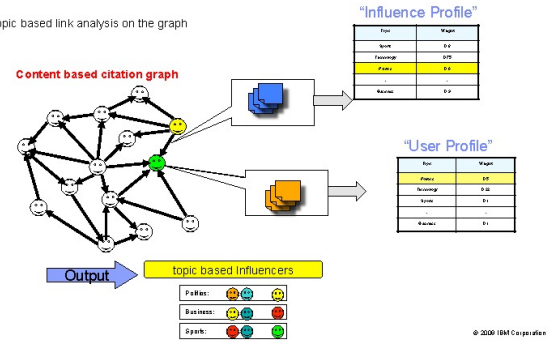


➢Precision over popularity improved by 28% at 1, 19% at 5

IBM

Finding Influencers in Social Media

- Build user profiles and "link" profiles (Blogs, Twitter…)
- Topic based link analysis on the graph



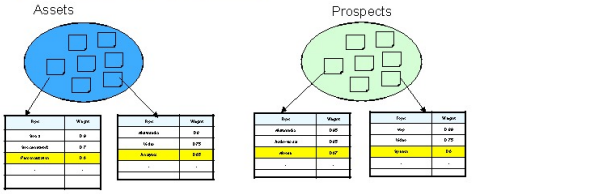
IBM

Similarity

➢Example: rank prospects by decreasing similarity to assets

➢Use assets and prospects profiles
➢Apply cos-similarity on each (a,p) pair
➢Take for each asset the top-k prospects

$$score(a,p) = \sum_{t \in profile(a) \cap profile(p)} w(t,a) * w(t,p) * idf(t)$$



IBM

Summary

- Using massive scale analytics to improve IBM customers business scenarios
  - analyzing more data (e.g., online behavior)
  - analyzing new types of data (e.g., text, social relationships)
  - developing finer grained models (personal models)

ARISTOTE Association Loi de 1901. Siège social : CEA-DSI CEN Saclay Bât. 474, 91191 Gif-sur-Yvette Cedex.
Secrétariat : Aristote, École Polytechnique, 91128 Palaiseau Cedex.
Tél. : +33(0)1 69 33 99 66 Fax : +33(0)1 69 33 99 67 Courriel : Marie.Tetard@polytechnique.edu
Site internet http://www.association-aristote.fr