Impact des architectures matérielles Exascale sur les environnements systèmes de calcul

05-02-2015

Atos

Quelles architectures pour les simulations de demain?:

- Une double question
  - Les besoins des simulations de demain
    - Evolution liée à la taille de la simulation et de ces données
    - Evolution liée à la décomposition et/ou capacité de distribution
  - les propriétés des architectures de demain
    - Contraintes d'Energie et cout
    - Contraintes de Qualité de service

Les besoins des simulations ? « Pas de besoin unique »

- Evolution des modèles d'exécutions
  - dimension des modèles
    - Vérification en cours de simulation (Traitement in situ)
    - Evolution du schéma d'accès aux données
    - Implémentation des points de reprise
    - Aide à la mise au point
    - Nouvelles modèles et langages de programmations (work stealing, graphs, …)
  - profil applicatif /plateforme
- Diversifications des besoins
  - Nouveaux types d'applications
    - Ex « Big Data » et Traitement sur des données non structurées (data analytics)
  - Phases de simulations
    - Pre-/post processing , visualisation etc …

Les défis des architectures matérielles et systèmes de demain 1/2

- Une performance dans une enveloppe énergétique. 20 megawatts ?
  - Introduction de nouvelles technologies « Low power» ,
    - unité d'éxécution hétérogène ,mémoire multi-niveaux
      - Xeon phi, ARM, GPU,
    - mémoires non volatiles, stackée, (NVRAM, eSSD…)
  - Définition du « Watts/flops » en production
    - Watts/flops applicatifs (Nœuds+Réseaux+stockages)
    - opérations matériels versus logiciels

- La robustesse et maintenabilité de l'infrastructure de calcul
  - Cablage, topologie et routage de très grandes fabriques
  - Optimiser le traitements d'erreurs
    - Propagations sélectives des erreurs
    - Confinements ou corrections au plus tot

Les défis des architectures matérielles et systèmes de demain 2/2

- La scalabilité des architectures matérielles
  - Optimisations du nombre de Nœuds => nœuds larges ou hybrides
    - Hyperthreading , NUMA
    - Mémoires multi-niveaux, affinités, cohérences
  - Optimisation des opérations d'échanges de données (Op collectives matérielles ,DMA , Zero copies)
  - Optimisation des Architectures I/O
    Support des checkpoints (local/global)
    Passage a l'échelle: Nombre de clients, Metadonnées, Nombre fichiers
- L'hétérogénéité des ressources
  - Nœuds d'IO, Nœuds Visualisation etc ….

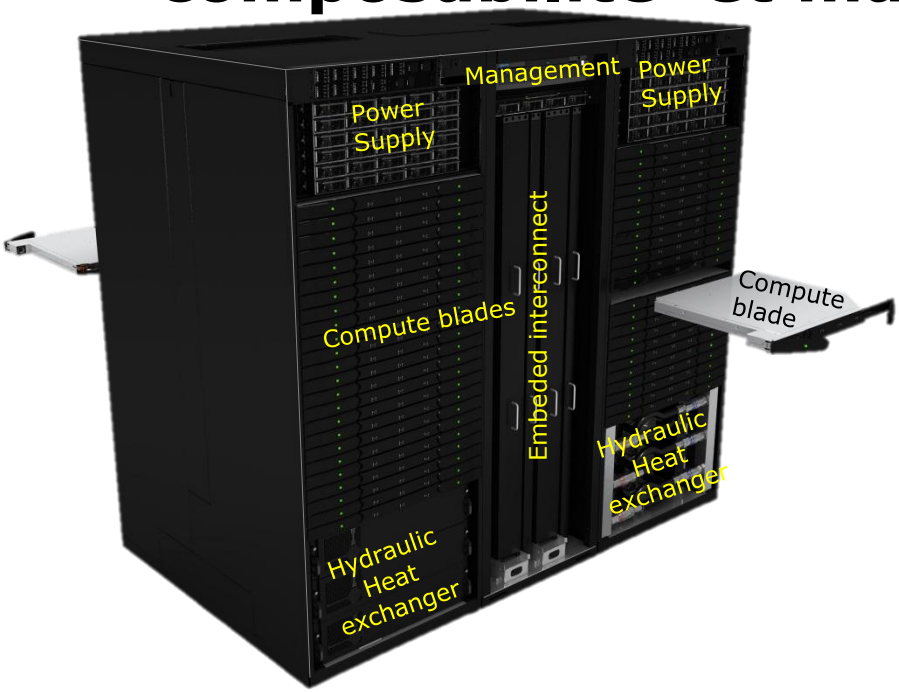
Programme « Bull »/Atos Exascale

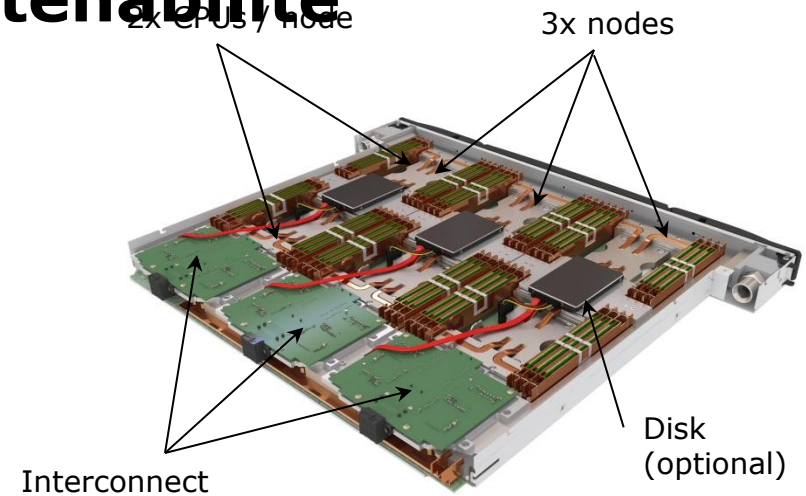
- Un programme exascale R&D (2015-2020) partenariat CEA
  - **Volet matériel**
    - **Packaging SEQUANA (modularité, maintenabilité)**
    - **Nœud Large (ex pre-post processing)**
    - **Interconnect BXI optimisations des opérations**
    - **Répartition des accès aux espaces de données (de NVM au stockage)**
  - Un volet logiciel
    performance & scalabilité,power, résilience
    Composabilité des environnements d'éxecution

- Des interactions avec les communautés

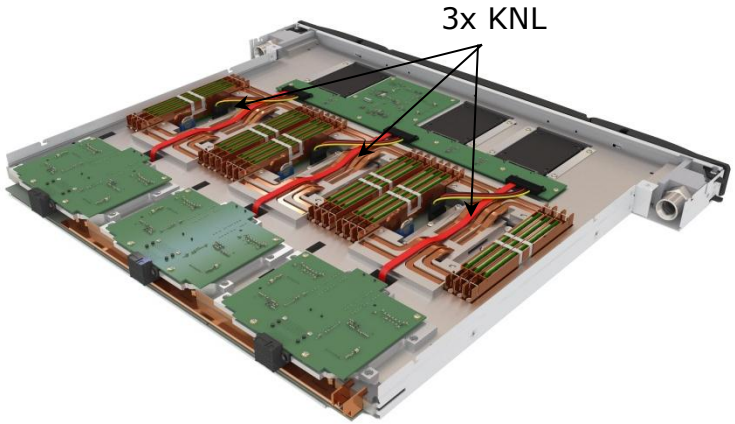
- Un centre d'excellence
  - accompagnement des développements applicatifs

Bull eXascale Platform “Sequana”: composabilité et maintenabilité





Xeon-EP CPU blade



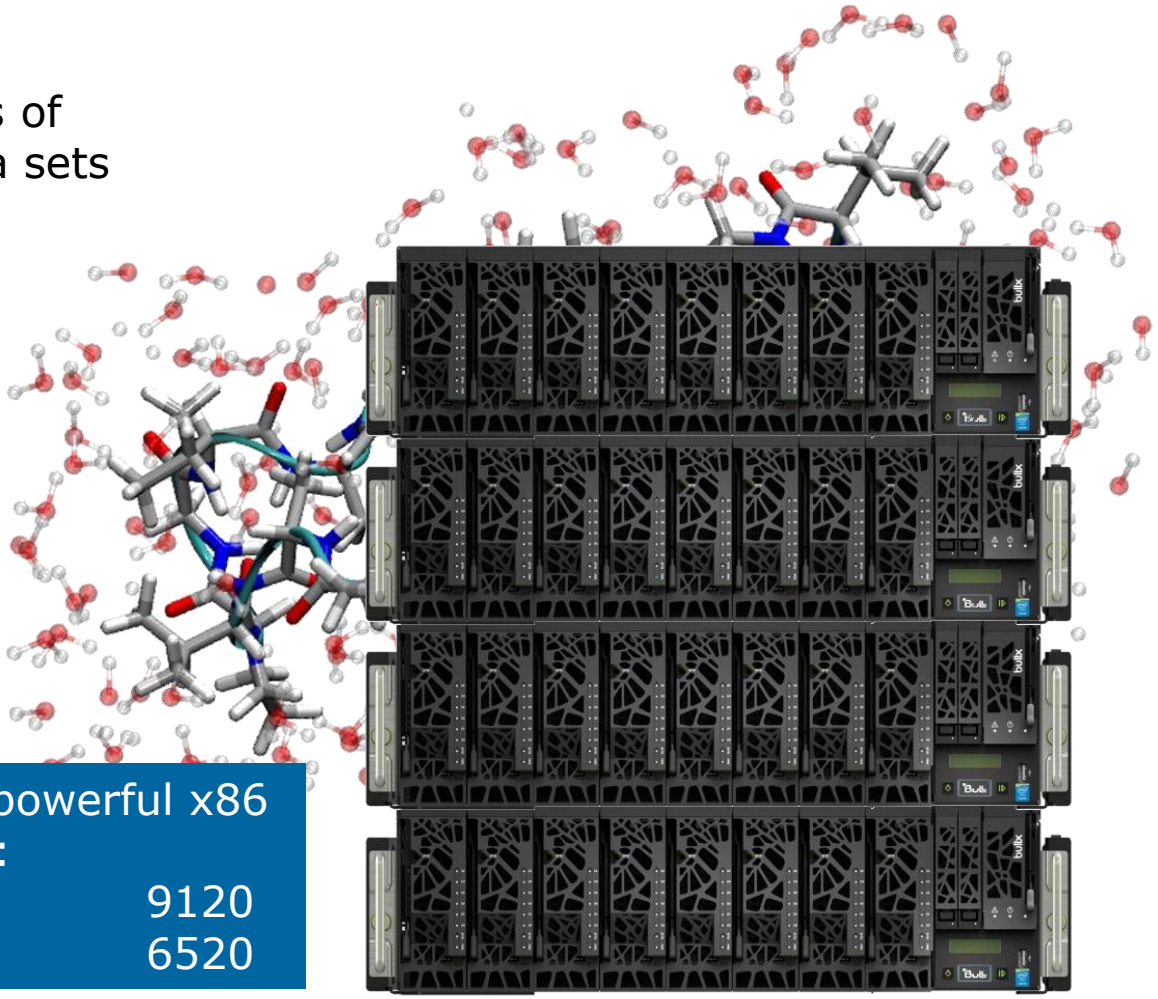
Xeon-PHI (KNL) blade

- Large building block: 288 nodes
- Direct liquid cooling
- 96 blades
  - **CPU blades – 3 nodes/blade**
  - **Xeon Phi blades**
  - **GPU blades**

The extra-large memory super computer

bullx S6000 series

- Fast compute and analysis of complex and massive data sets
- Designed for
  - In-memory data bases
  - Pre-processing
  - Post-processing
  - Visualization
- Up to 24 TB
- Up to 16 sockets



HPC version of the #1 most powerful x86 enterprise server bullion s16:

SPECint_rate2006	9120
SPECfp_rate2006	6520

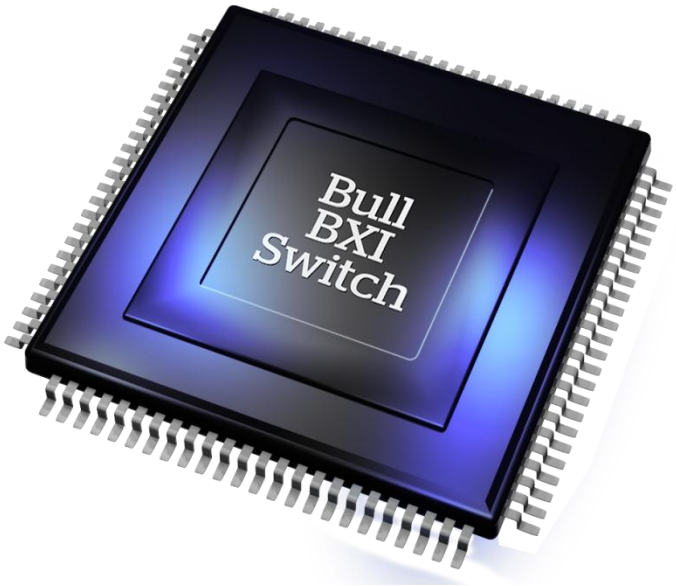
BXI, the fastest interconnect for exascale (Bull Interconnect for the Exascale)

- BXI frees up CPU performance
  - Hardware acceleration to eliminate communication overhead

- High-bandwidth
- Low latency
- High message rate
- Multi data-flows to maximize quality of service

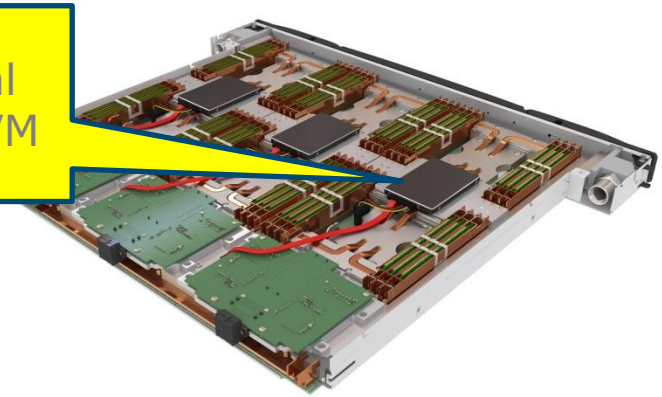
- Cutting-edge technologies
  - 100 Gb/s (4X25)

- Scaling to tens of thousands of nodes

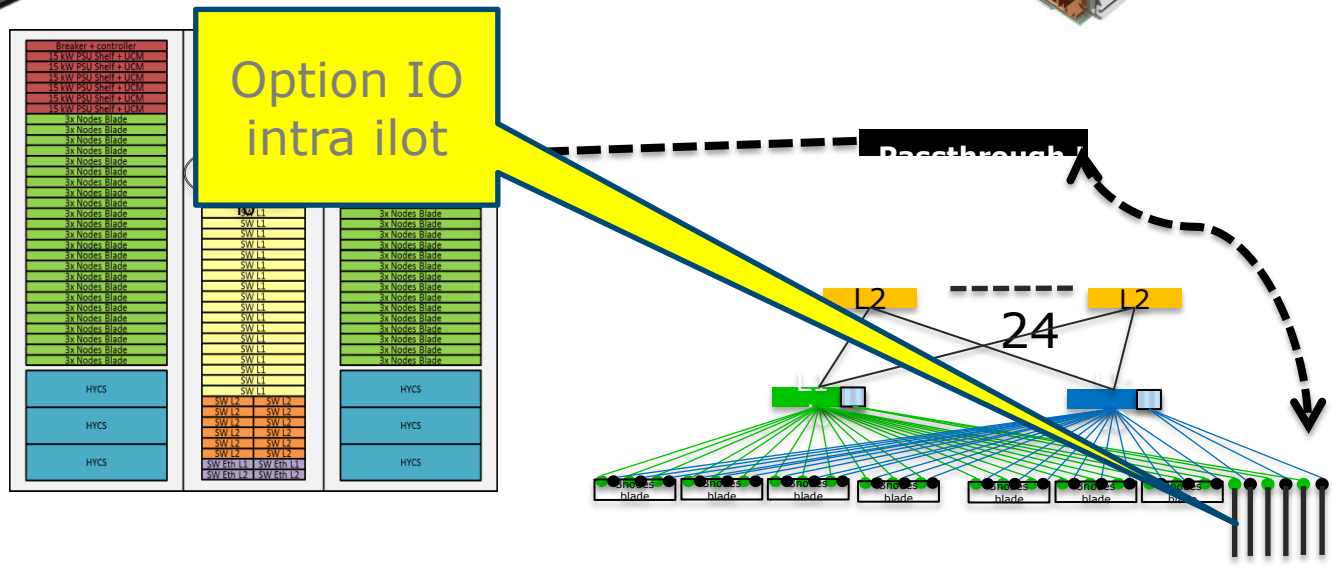


Support et localisation des espaces de données

Option IO local
Flash, essd, NVM



Option IO
intra ilot



Programme « Bull »/Atos Exascale

- Un programme exascale R&D (2015-2020) partenariat CEA
  - Volet matériel
  - **Volet logiciel**
    performance & scalabilité, power, résilience….
    - **Administration**
    - **systèmes et réseaux optimisées**
    - **Environnements d'exécutions et supports d'executions**
    - **IO et les nouveaux espaces de données**

- Des interactions avec les communautés

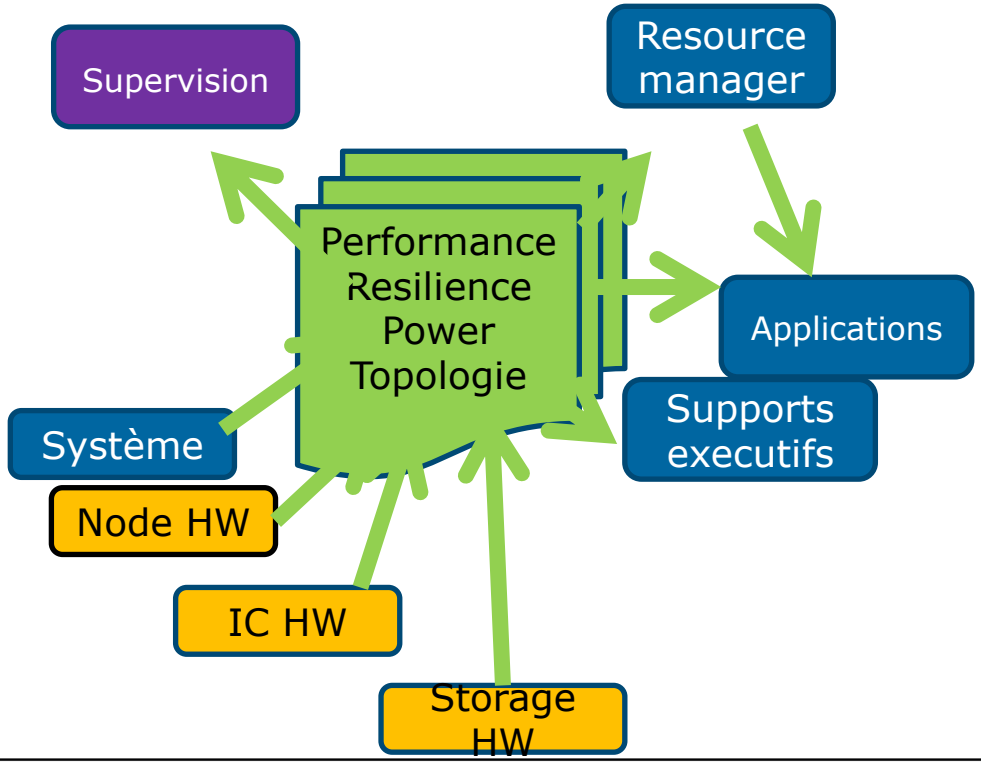
- Un centre d'excellence accompagné les développements applicatifs

Les données « d'infrastructure ou système » une mine d'information à partager

- Une masse de Données
  - une architecture modulaire et scalable d'administration
  - Des espaces de données

- Une mine d'information à partager
  - **Des formats à standardiser**
  - **Des API à spécifier**

- Exploitations
  - in situ
    - Placement /optimisation
    - Politique de correction ou reprise sur erreur
  - Post-mortem
    - Profilling , data mining ….



Environnements d'exécution

- Piles d'exécution
  - Optimisation de modèles hybrides, PGAs , ou futurs (threads ,graphes , taches …)
  - Prise en compte de la **topologie** mémoires, cœurs, réseaux
  - Dynamicité, interopérabilité des composants logiciels
- Passage à l'échelle des outils
  - Méthode de lancement , espaces de données
- Management Resource multi critères , ressources hétérogènes (slurm)
- Supports d'exécutions
  - Adaptation de l'exécution à la diversification des ressources de calcul
  - Prise en compte de la topologie des ressources

Architecture IO et espaces de données

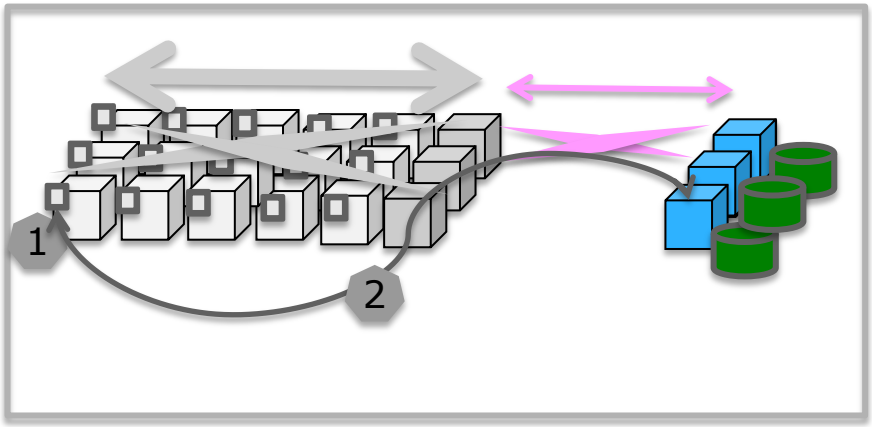
- Grand système de stockage et système de fichiers global trop lent / calcul

- Améliorer certaines opérations de lecture/écriture critiques
  - Utilisation des technologies ultra rapides: NVM, Flash ou appliance Hybrid Flash-SSD
  - Développement de pile software adaptée et standardiser les interfaces

- Principaux besoins
  - Support des checkpoints
  - Caching IO entre nœuds de calcul et infrastructure de stockage « lente »
  - Pre-post et in situ processing pour certains type d'applications

Espaces et fonctionnalités

- Checkpoints « locaux »

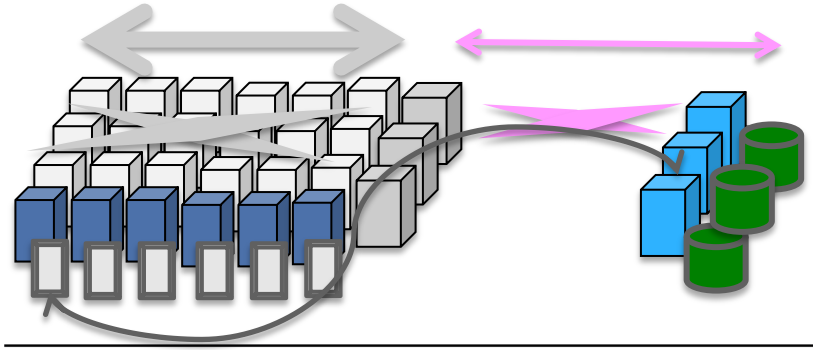


1 Ecriture local 400MB/s
2 Asynchroneous slow R/W to Global FS

Cons: Local Daemon for Ckpt mgt
Not useable as global IO cache

Pro: st buffer Global management
Limited global bandwith need for Asynch R/W From/to FS

- Visualisation et Pre-post processing



1 Nœud Calcul spécialisé Visu ou pre-post processing
2 Asynchroneous slow R/W to Global FS

Cons: Nombre Fixe ,
gérés comme un pool de ressource dédiées
pas utilisables en tant que cache IO …

Pro: utilisations de nœuds larges
Limited global bandwith need for Asynch R/W From/to FS

Programme « Bull »/Atos Exascale

- Un programme exascale R&D (2015-2020) en partenariat avec le CEA
  - Volet matériel
  - Volet logiciel

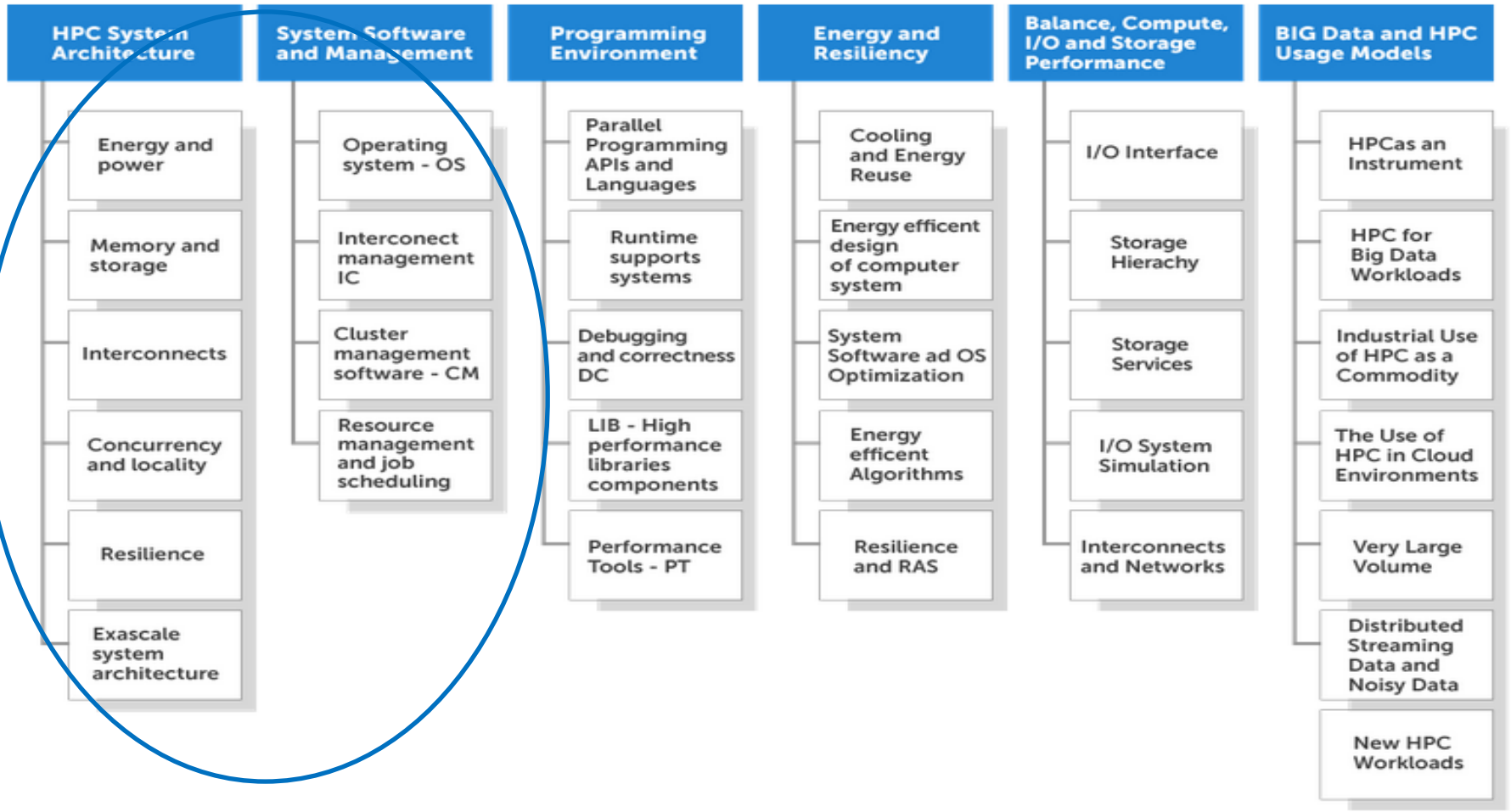
- Des interactions ou participations avec les communautés
  - **Communautés open sources (slurm, lustre, openmpi….)**
  - **Communautés exascale (EESI,ETP….)**
  - **Programme de recherche (MOEBUS,COLOC, ELCI, DATASCALE)**
  - **PCP et H2020**

- Un centre d'excellence en programmation parallèle
  - Accompagnement des projets et développements applicatifs

ETP for HPC : Strategic Research Agenda

RESEARCH TOPICS

Through the SRA, the ETP4HPC has identified research areas and topics which were considered priorities to reach the objective of a stronger European HPC environment that can benefit Europe and the rest of the world.



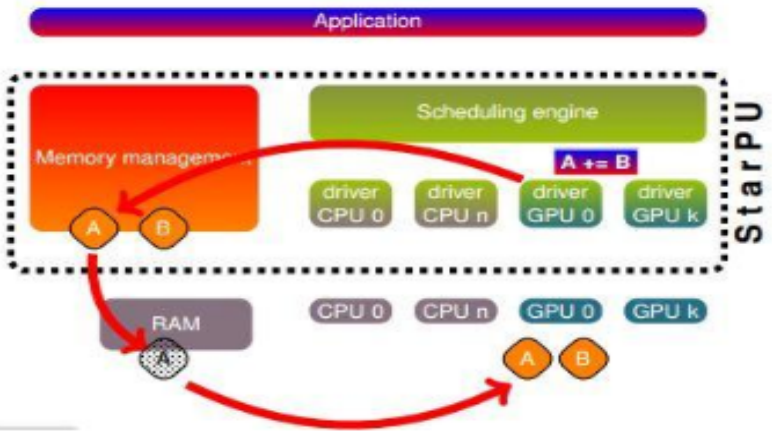
Projets de recherche pour préparer l'exascale

- COLOC:
  - Integration de la localité dans les politiques de placement
- DATASCALE:
  - Optimisation du placement de données
- MOEBUS :
  - Gestion de ressources et scheduling multi-objective (power, type de ressources , …)
- HDEEM:
  - Sonde FPGA pour une mesure fine de la consommation électrique
- ELCI:
  - Preuve de concept d'une pile logicielle optimisée pour des grands calculateurs
    - solveurs numériques et bibliothèques optimisés et performants et profitant pleinement du calculateur HPC (hétérogénéité, résilience, scalable….)
    - des cas d'utilisation concrets (domaine de l'aéronautique, domaine de la fabrication des procédés).

ELCI: Interaction système/run time et applications

T1.4: Supports d'exécution

- Objectif: Faire que les politiques d'exécution soient prises en charge par les runtimes MPI, PGAS, MPC, OpenGL, Xkaapi, StarPU et non plus au niveau applicatif et ceci dans le but d'une adaptation dynamique aux ressources matérielles en cours d'exécution.
- Spécification des info pertinentes (topo, compteurs hard, format d'err) et modes de mise à dispo (Bull)
- Spécification d'un nouveau standard (INRIA)
- Etude des différentes possibilités d'extension MPI (INRIA)
- Mise à dispo d'interfaces prototype et n
  en œuvre sur un exemple (Bull)
- Implémentation et standardisation MPI forum (INRIA)

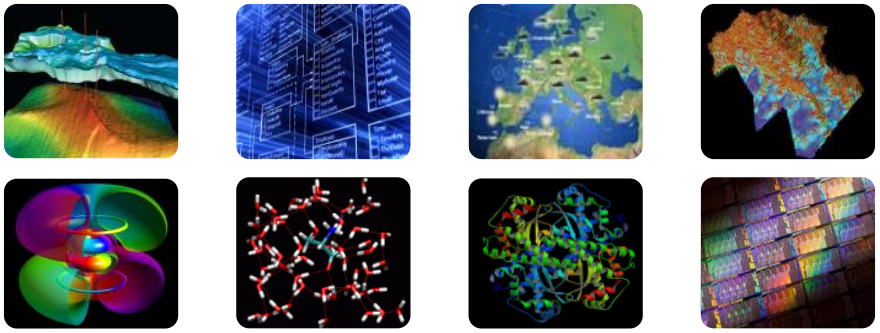


Programme « BULL »/Atos Exascale

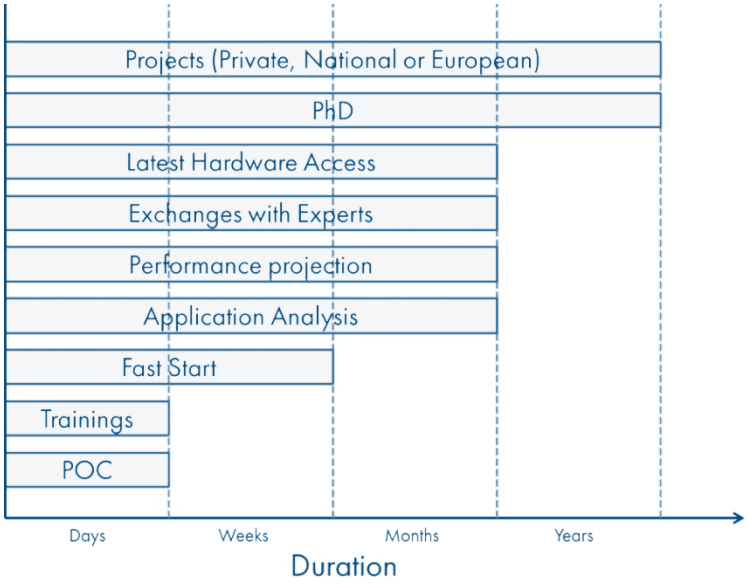
- Un programme exascale R&D (2015-2020) en partenariat avec le CEA
  - Volet matériel
  - Volet logiciel
- Des interactions ou participations avec les communautés
- **Un centre d'excellence en programmation parallèle**
  - **Accompagnement des projets et développements applicatifs**

Préparer les applications pour l'exascale

- L' Exascale est un défi pour les applications
  - Architecture materiel
  - Concurrence d'accès aux données et ressources
  - Efficacité énergétique



- Le centre d'excellence en programmation parallele
  - Accompagnement de la re-ecriture des applications
  - Optimisation pour les architectures exascale
  - Resilience
  - Etc …



Thanks

For more information please contact:
T+ 33 4 76297078
Pascale.rosse-laurent@atos.net

05-02-2015