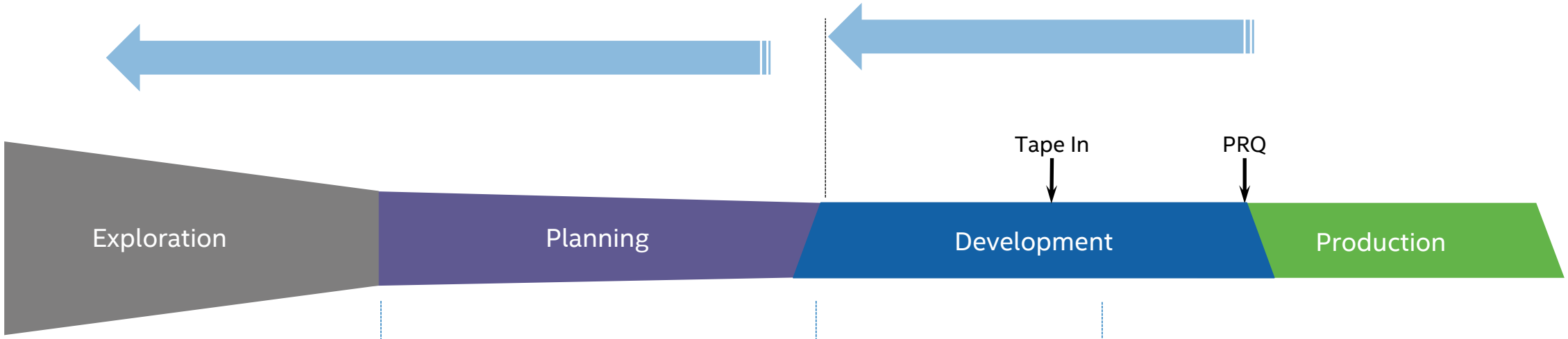
Challenges in performance modelling at scale

intel AI-Accelerated HPC

Process



Design

u-Arch Functional simulator
u-Arch Performance simulator

Kernel compute prediction
+ Application communication prediction

Real hardware measurements

Apps-level

Full application shared memory (i.e. node level)

Full application distributed memory

Full application hybrid memory (full system, full MPI + OMP application)

iterate

Real hardware measurements

Applications : all about communications and compute

Any E2E modeling is (or should be) doing **comms** ⊕ **compute** Where ⊕ can be either

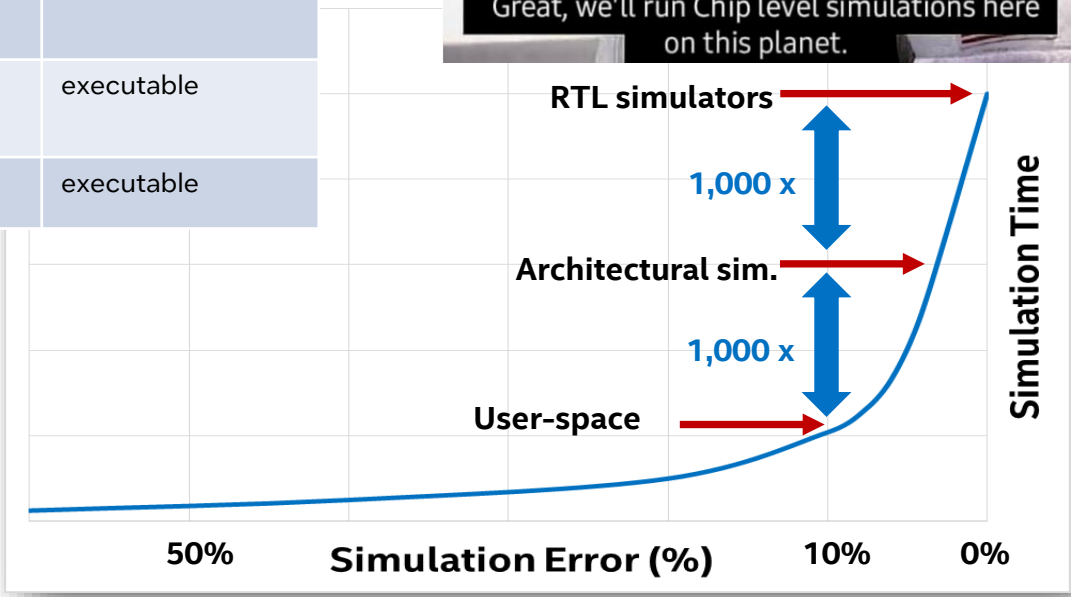
- **An addition** : not accounting for overlap, asynchronism
- **A more complex** “tracing and replay” to preserve dependencies

	Communications	Compute
tracing	Tracing, aggregation	Papi, Advisor, Emon
modeling	LogP , SIMGRID, SST…	Analytical , Cycle Accurate….

- Scalability: how easy to move to any other applications.

Simulation landscape for “shared memory” systems

Simulator	Purpose.	Error	Scale	Speed	Input
cycle-accurate RTL simulators	• design verification	0%	1	1 I/sec	traces
near-cycle-accurate simulators	• micro-architectural exploration and evaluation • performance projections (single core)	~2-5%	1e3	1 k.I/sec	traces
User-space	• architectural exploration • performance projections • application, compiler, library optimization	~5-10%	1e6	1 M.I/sec	executable or traces
User-space	• functional only • user-space only	no performance results	1e9	1 G.I/sec	executable
Applications	• Real runs	n.A	1e12,15,18	T, P , E/sec	executable



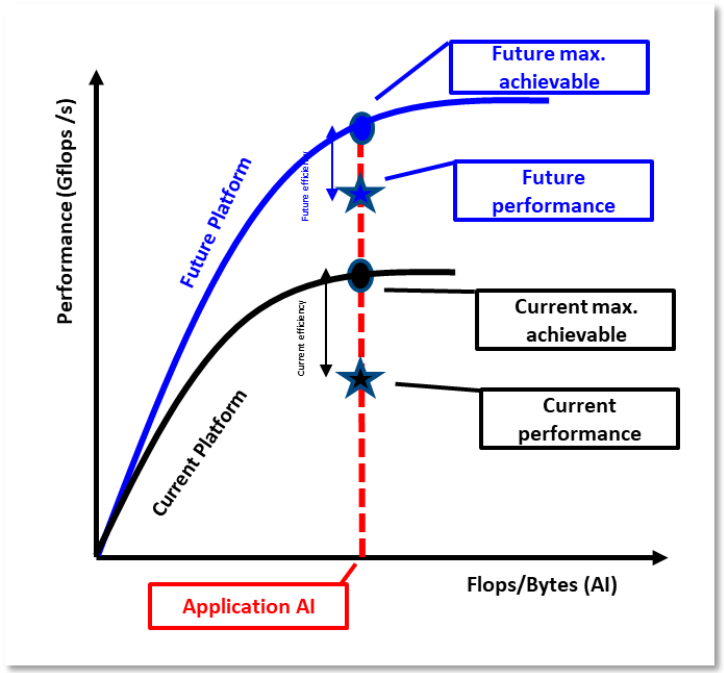
Accuracy and speed are orthogonal features

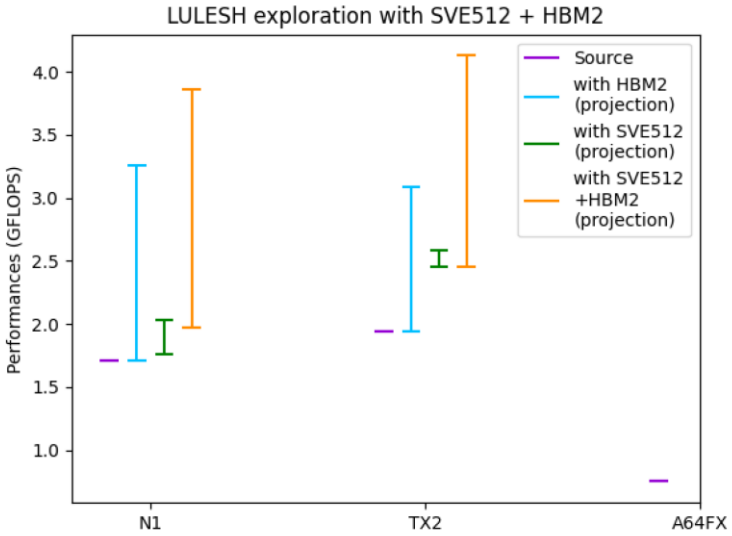
1mn of a Real application = 20y with cycle accuracy

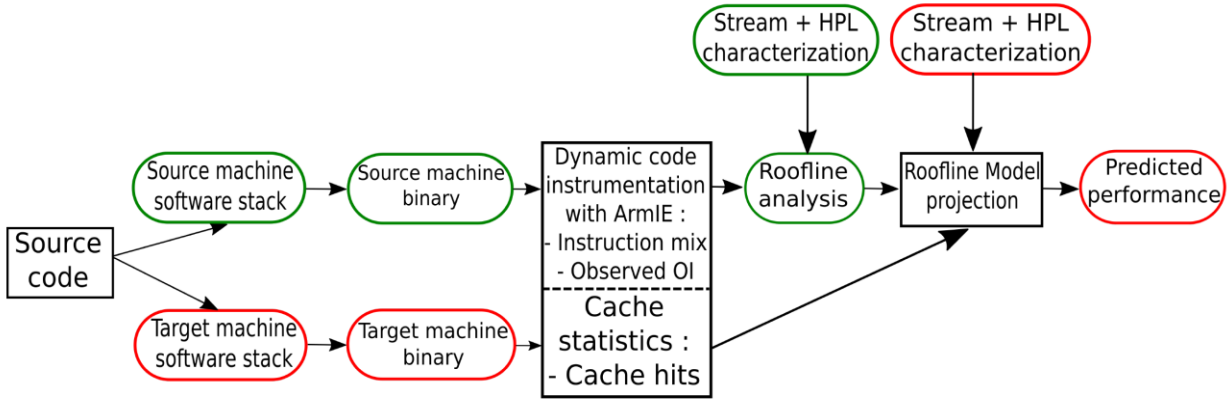
intel AI-Accelerated HPC

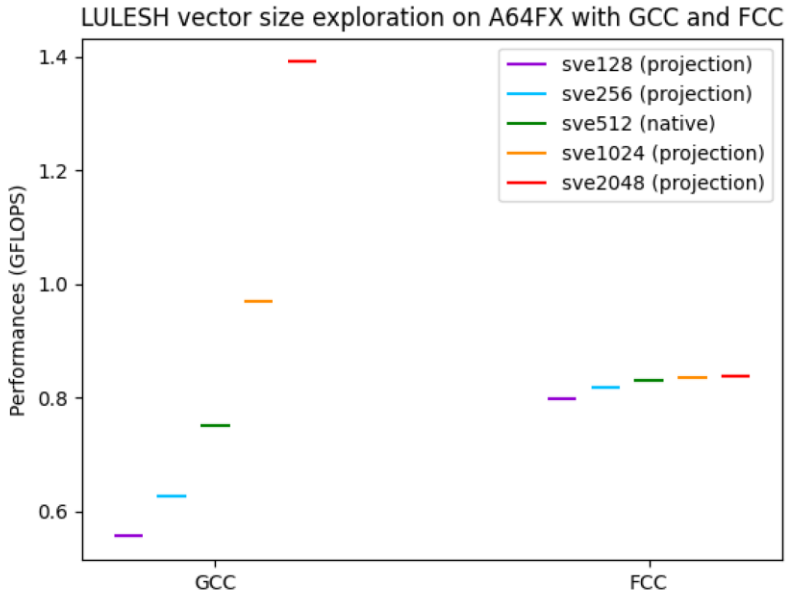
“What if” Analyses

Gavoille et al. 2022





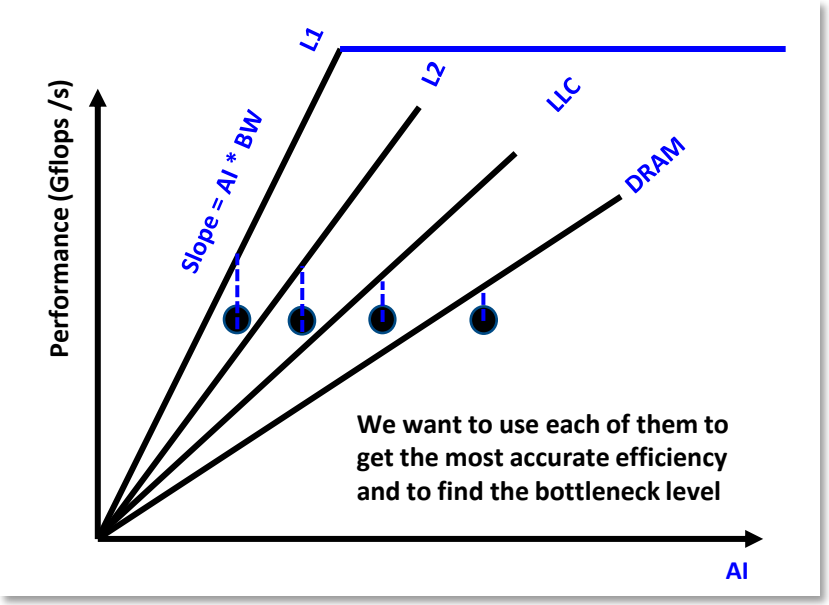




“What if” Analyses

Offload Advisor

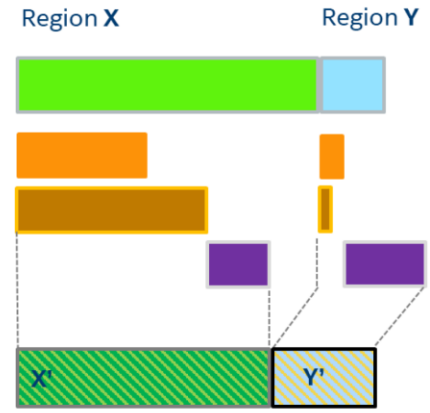
- Portion of code profitable to be offloaded
- Predict code’s performance if run on an accelerator



Execution time on baseline platform (CPU)

- Execution time on accelerator. Estimate assuming bound exclusively by Compute
- Execution time on accelerator. Estimate assuming bound exclusively by caches/memory
- Offload Tax estimate (data transfer + invoke)

Final estimated time on target device (GPU)

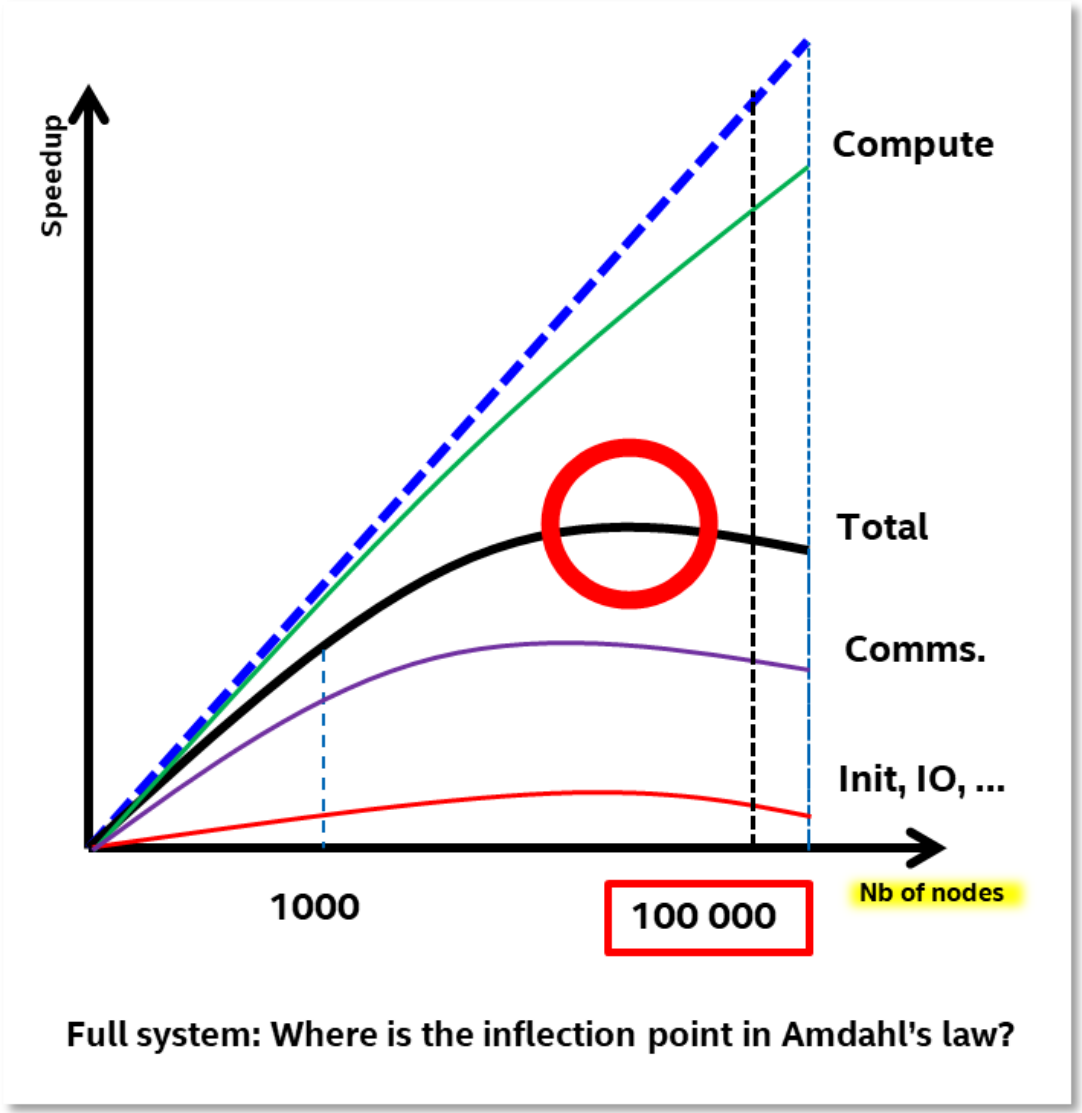


X – profitable to accelerate, $t(X) > t(X')$

Y - too much overhead, not accelerable, $t(Y) < t(Y')$

$$t_{region} = \max(t_{compute}, t_{memory\ subsystem}) + t_{data\ transfer\ tax} + t_{invocation\ tax}$$

Simulation landscape at Scale



Structural Simulation Toolkit (SST)

- Scalable, parallel DES* aimed to simulate large-scale systems
  - Developed by Sandia National Laboratories (http://sst-simulator.org/)
- Key Features
  - Scalability
    - Uses MPI for graph partitioning to enable million + component simulation
  - Multi-scale
    - Provides both simple and detailed (~ cycle) models for memory, compute and network components
  - Inter-operability
    - Provides support to integrate third-party component models
      - E.g., gem5, GPGPUSim etc..
  - Open-source, modular and easily extendable
- Key Components
  - SST-Core: Simulation Backend
    - Takes care of simulation startup, component partitioning, event passing etc..
  - SST-Elements: Library containing detailed component models
    - E.g.: Merlin (router), DRAMSim2 (memory), Ariel (CPU)

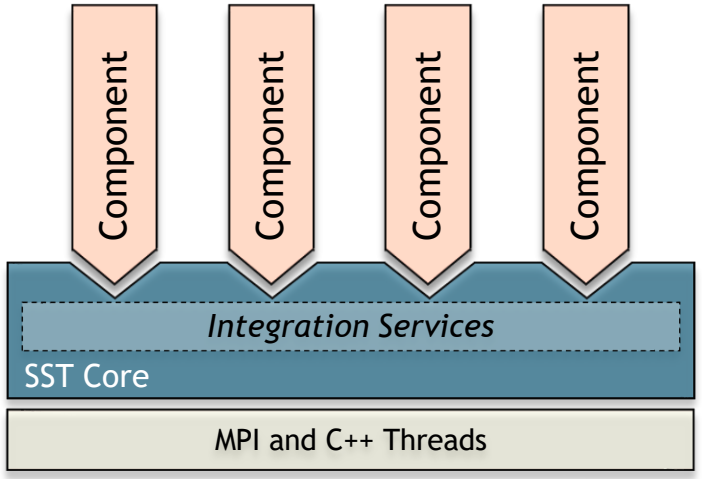
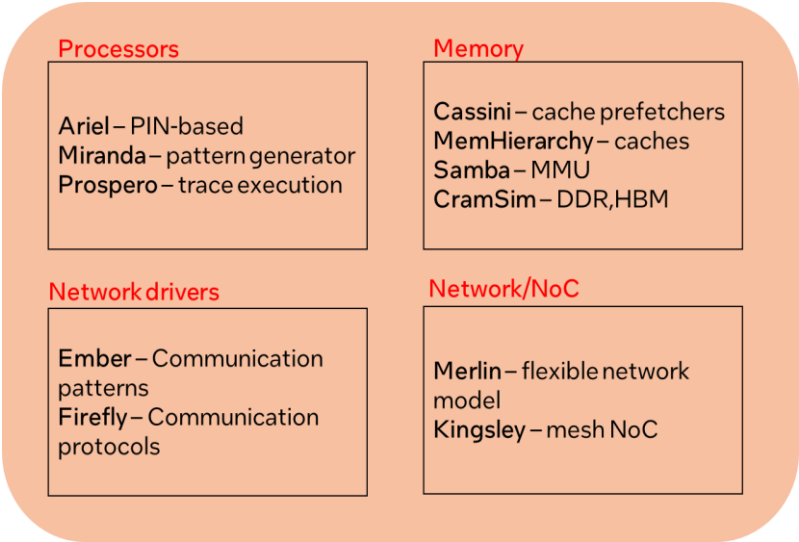
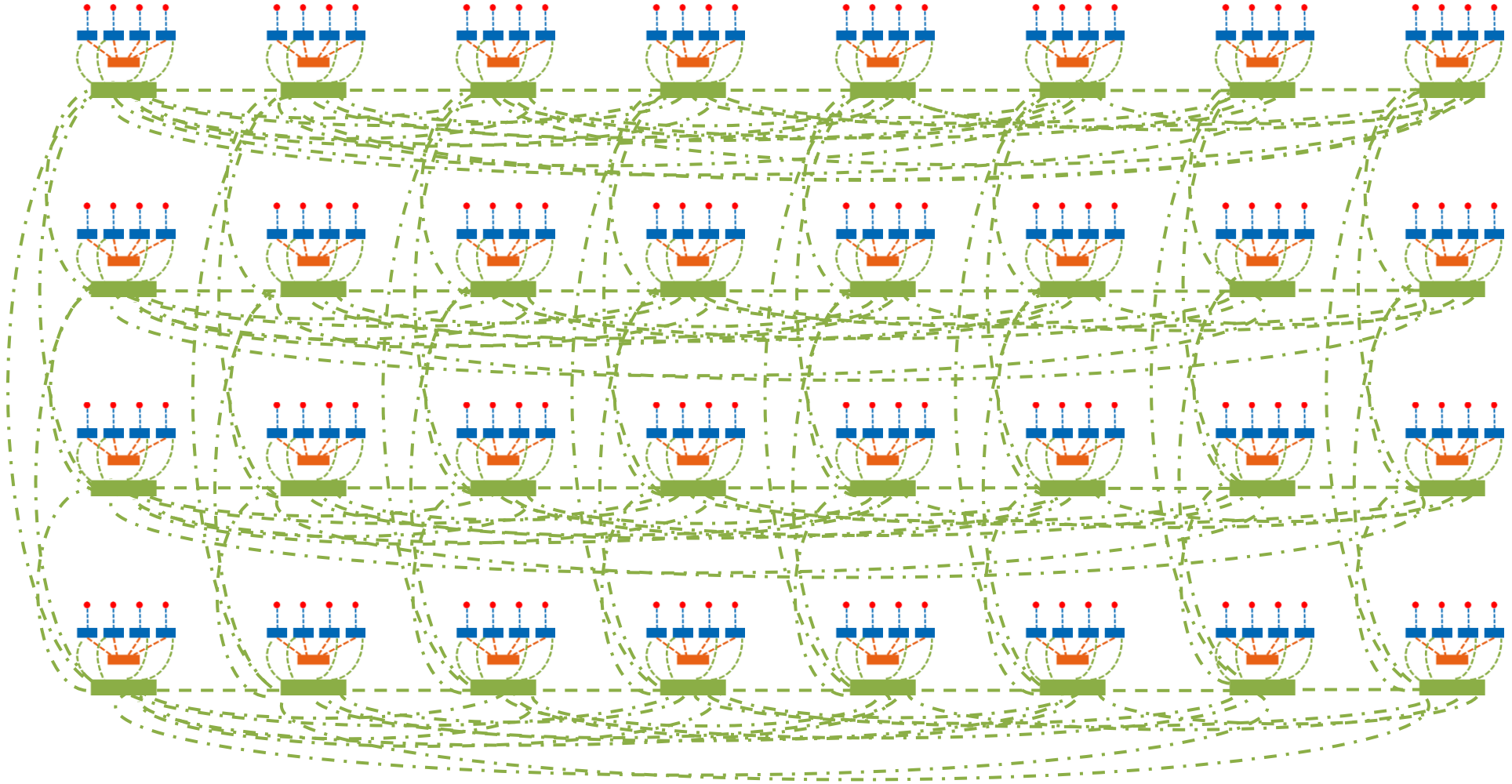


Image courtesy: Sandia National Laboratories (http://sst-simulator.org/SSTPages/SSTTopDocTutorial/)



Modeling Multi-level Networks: Two-level Network

Sai et al. 2022



Intra-node: Fully connected
Inter-node: 2-D HyperX (8,4)
Node-size: 4
Node-count: 32

endpoint
endpoint NIC
intra-node switch
inter-node switch

Two-level network with 128 endpoints

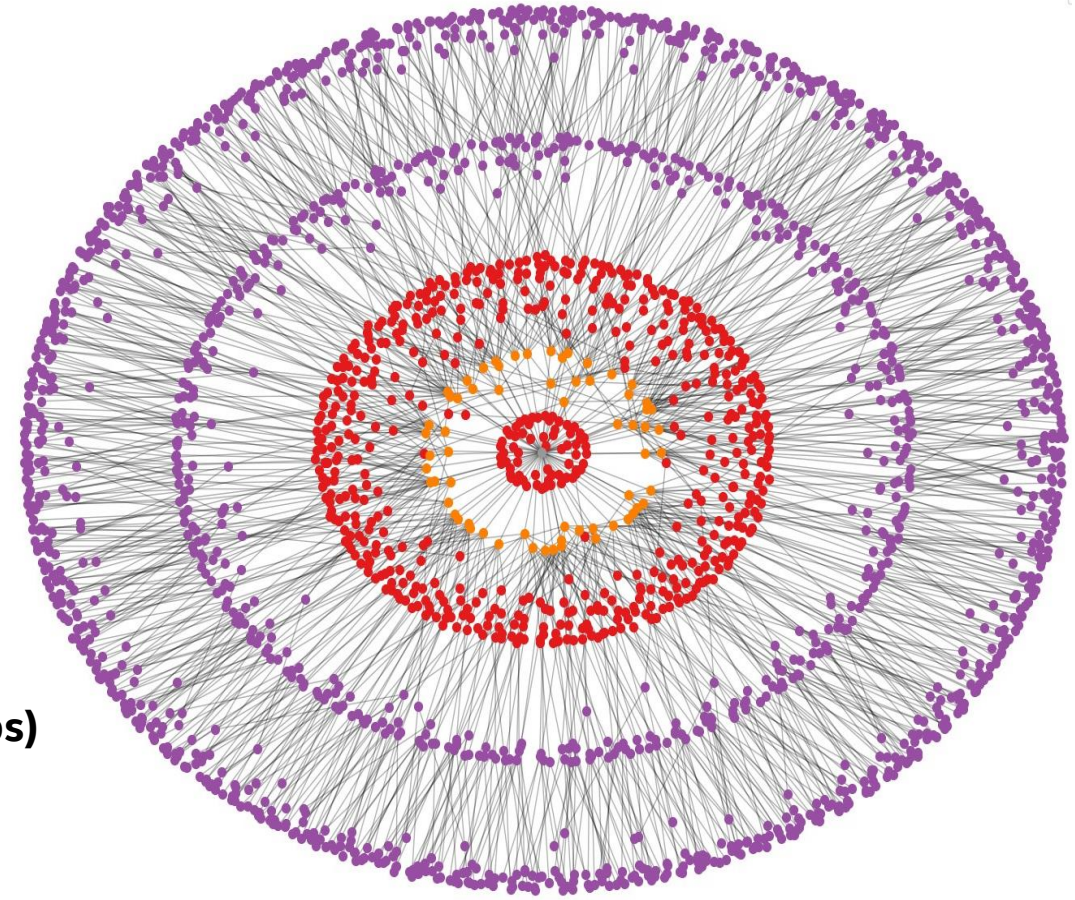
intel AI-Accelerated HPC

SIMGRID - Simulation of distributed systems

Simulation of distributed systems

- Leveraging Two decades of validation and improvement driven by Inria + collabs (https://simgrid.org/)
- 200+ papers
- Open source

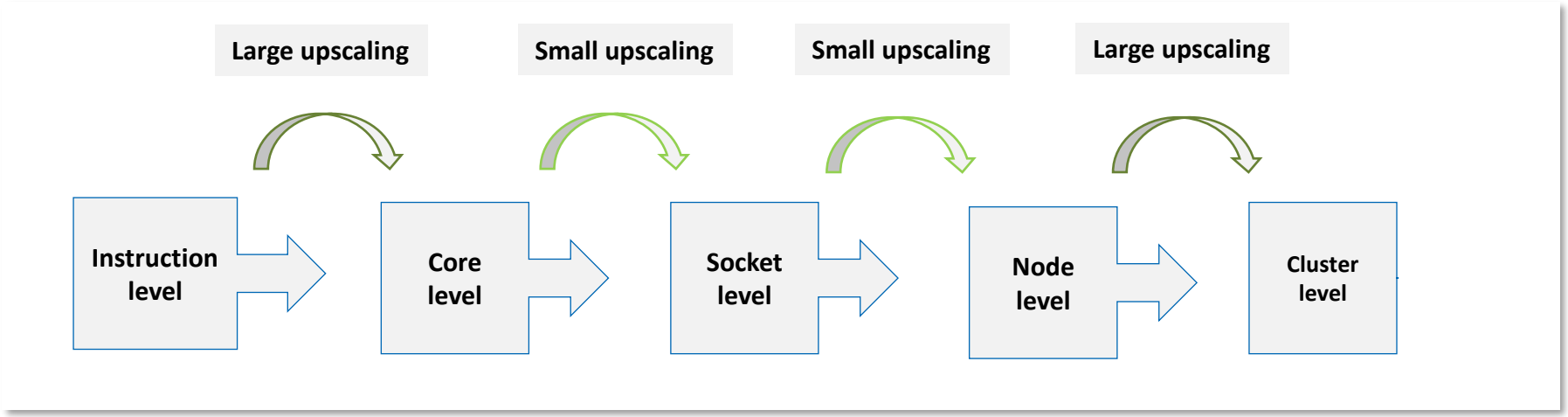
- Platforms (machines, networks, disks ...)
- Topology (fat-tree/dragonfly/cluster ...) including congestion.
- MPI operations : multiple algos/implementations available
- Modelling of both compute and communication phases



- Online mode
- Leveraging PMPI interface to intercept MPI calls **(Unmodified apps)**
- Offline mode
- Generate / Replay traces

Casanova et al. 2014

A multi-scale problem



1. Granularity is extremely different from instruction up to cluster levels
2. Merging different tools that have different accuracy and speed
3. Uncertainties to validate final prediction accuracy
4. Relevant benchmarks suite is key

- Ongoing efforts


Thank You!
intel
AI-Accelerated
HPC