SAND2012-5445C

Applications Programming at Exascale
– Challenges and Opportunities –

Robert L. Clay, Ph.D.
Sandia National Laboratories

ISPA 2012 Keynote
July 11, 2012
Madrid, Spain

Sandia National Laboratories is a multi-program laboratory operated by Sandia Corporation, a wholly owned subsidiary of Lockheed Martin company, for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-AC04-94AL85000.

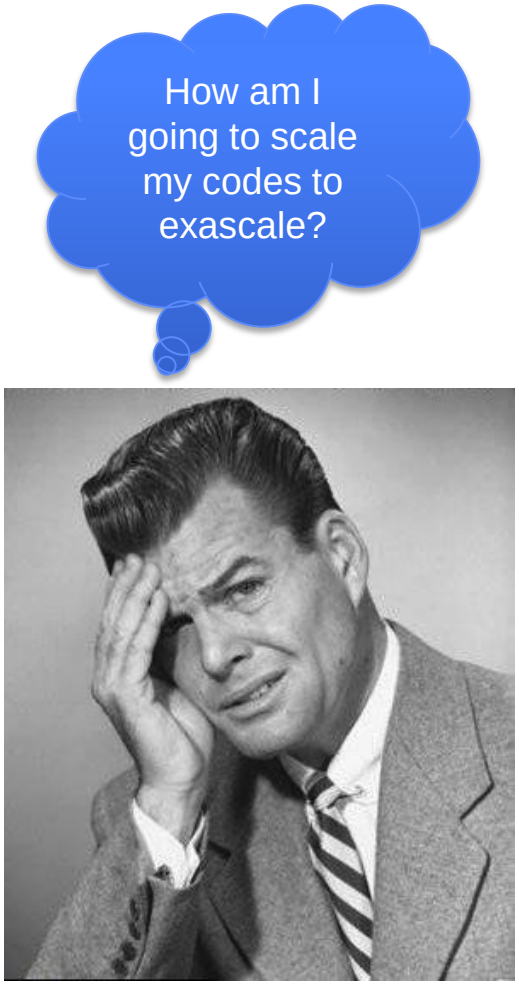
Sandia National Laboratories

Exascale Changes Everything

Robert L. Clay, ISPA-12

Sandia National Laboratories

And those changes present serious challenges to apps developers



Outline

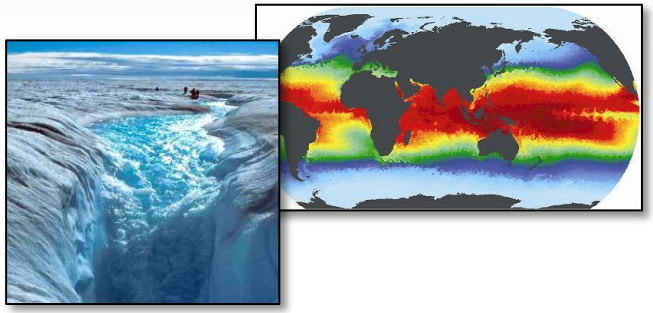
- **Applications**
- **Trends**
- **Challenges**
- **Selected work at Sandia**
- **Preparing for exascale**

APPLICATIONS

Science and engineering are the real drivers.

DOE mission imperatives require simulation and analysis for policy and decision making

- ***Climate Change*: Understanding, mitigating and adapting to the effects of global warming**
  - **Sea level rise**
  - **Severe weather**
  - **Regional climate change**
  - **Geologic carbon sequestration**
- ***Energy*: Reducing U.S. reliance on foreign energy sources and reducing the carbon footprint of energy production**
  - **Reducing time and cost of reactor design and deployment**
  - **Improving the efficiency of combustion energy systems**
- ***National Nuclear Security*: Maintaining a safe, secure and reliable nuclear stockpile**
  - **Stockpile certification**
  - **Predictive scientific challenges**



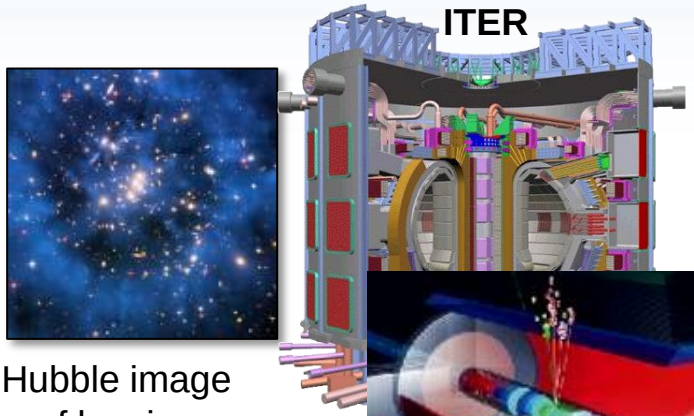
Accomplishing these missions requires exascale resources.

Courtesy of Sudip Dosanjh

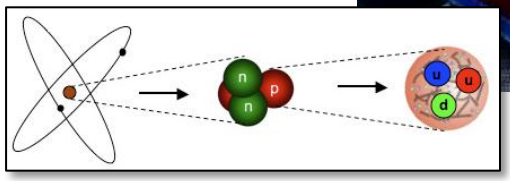
Exascale simulation will enable fundamental advances in basic science

- **High Energy & Nuclear Physics**
  - **Dark-energy and dark matter**
  - **Fundamentals of fission fusion reactions**
- **Facility and experimental design**
  - **Effective design of accelerators**
  - **Probes of dark energy and dark matter**
  - **ITER shot planning and device control**
- **Materials / Chemistry**
  - **Predictive multi-scale materials modeling: observation to control**
  - **Effective, commercial technologies in renewable energy, catalysts, batteries and combustion**
- **Life Sciences**
  - **Better biofuels**
  - **Sequence to structure to function**

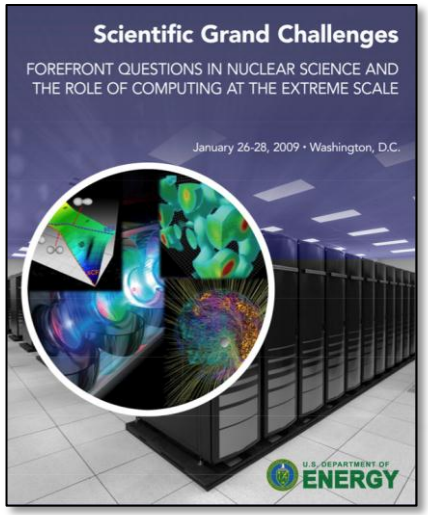
These breakthrough scientific discoveries and facilities require exascale applications and resources.

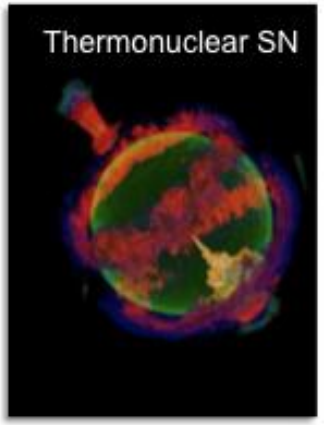


Hubble image of lensing



Structure of nucleons





Sandia National Laboratories

Courtesy of Sudip Dosanjh

TRENDS

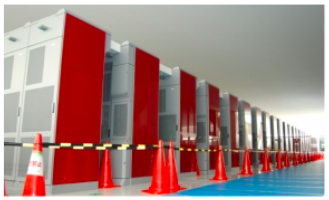
Exascale computing will happen.

IT'S INTERNATIONAL

China, Europe, India and Japan are making significant investments in supercomputing

Japan To Invest $1.3 Billion In New Supercomputer

SERKAN TOTO | posted on Friday, August 12th, 2011 | 5 Comments



There is a list of the world's 500 m
supercomputers, and the last tim
updated, back in June this year, F
(pictured) came out on top, taking
from Tianhe-1A (a supercomputer

It was the first time since 200
claim those bragging rights, a
largest business newspaper T
that the government is already
will happen in 2020: by then,
a computer that handles exascale computing or, in other words, one million trillion
second (that computer would be 100 times more powerful than K).

technology review

Published by MIT

Advertisement

SPECIAL REPORT: UNDERSTANDING THE CUSTOMER

Hitting it perfectly is anothe

English | en Español | auf Deutsch | in Italiano | 中文 | in India

HOME | COMPUTING | WEB | COMMUNICATIONS | ENERGY | MATERIALS | BIOMEDICINE | BUSINESS | MAGA

Enter China: A prototype four-core Loongson 3 will be produced at commercial scale by STMicro starting this year.
Credit: Institute of Computing Technology, Chinese Academy of Sciences

COMPUTING

China Details Homemade Supercomputer Plans

The machine will use an unfashionable chip design.

TUESDAY, JANUARY 19, 2010 | BY CHRISTOPHER MIMS

Audio »

It's official: China's next supercomputer, the petascale Dawning 6000, will be constructed exclusively with home-grown microprocessors. Weiwu Hu, chief architect of the Loongson (also known as "Godson") family of CPUs at the Institute of Computing Technology (ICT), a

Business

Home | Companies | Marke
Economy | Info-Tech | Agri-Biz

Plan panel to focus on adding to supercomputing capabilities

G. SRINIVASAN

SHARE · PRINT · T+

Dr Ashwani Kumar

NEW DELHI, MAY 25:

India is all set to step up its supercomputing capabilities and capacity with the Planning Commission seeking to take this as "a principal initiative in the Twelfth Five-Year Plan (2012-17) in consultation with other line Ministries and research and scientific institutions".

Disclosing this to *Business Line* here in an exclusive talk, the Minister of State for Planning and Science & Technology and Earth Science, Dr Ashwani Kumar, said that the Plan panel has got a specific report commissioned on the subject and recently flagged off the first round of discussion.

CHINA'S CAPABILITIES

Elaborating upon the "overarching" priority to underpin supercomputing capabilities of the country, Mr Kumar contended that in 2007-08, the country's supercomputing capabilities were more or less equal to that of China but since then

Manufacturing report

HPC in the Cloud

2011

January 06, 2011

European Exascale Project Drives Toward Next Supercomputing Milestone

With petascale systems now deployed on three continents, the HPC industry is already looking toward the next milestone in supercomputing: exascale computing.

Courtesy of Sudip Dosanjh

Sandia National Laboratories

In 2002 the United States Dominated Top 10

TOP 10 Sites for June 2002

For more information about the sites and systems in the list, click on the links or view the complete list.

Rank	Site	Computer
1	The Earth Simulator Center Japan	Earth-Simulator NEC
2	Lawrence Livermore National Laboratory United States	ASCI White, SP Power3 375 MHz IBM
3	Pittsburgh Supercomputing Center United States	AlphaServer SC45, 1 GHz Hewlett-Packard
4	Commissariat a l'Energie Atomique (CEA) France	AlphaServer SC45, 1 GHz Hewlett-Packard
5	NERSC/LBNL United States	SP Power3 375 MHz 16 way IBM
6	Los Alamos National Laboratory United States	AlphaServer SC45, 1 GHz Hewlett-Packard
7	Sandia National Laboratories United States	ASCI Red Intel
8	Oak Ridge National Laboratory United States	pSeries 690 Turbo 1.3GHz IBM
9	Lawrence Livermore National Laboratory United States	ASCI Blue-Pacific SST, IBM SP 604e IBM
10	IBM/US Army Research Laboratory (ARL) United States	pSeries 690 Turbo 1.3GHz IBM

In 2012 the Top 10 List is International

TOP 10 Sites for June 2012

For more information about the sites and systems in the list, click on the links or view the complete list.

Rank	Site	Computer
1	DOE/NNSA/LLNL United States	**Sequoia** - BlueGene/Q, Power BQC 16C 1.60 GHz, Custom IBM
2	RIKEN Advanced Institute for Computational Science (AICS) Japan	K computer, SPARC64 VIIIfx 2.0GHz, Tofu interconnect Fujitsu
3	DOE/SC/Argonne National Laboratory United States	**Mira** - BlueGene/Q, Power BQC 16C 1.60GHz, Custom IBM
4	Leibniz Rechenzentrum Germany	**SuperMUC** - iDataPlex DX360M4, Xeon E5-2680 8C 2.70GHz, Infiniband FDR IBM
5	National Supercomputing Center in Tianjin China	**Tianhe-1A** - NUDT YH MPP, Xeon X5670 6C 2.93 GHz, NVIDIA 2050 NUDT
6	DOE/SC/Oak Ridge National Laboratory United States	**Jaguar** - Cray XK6, Opteron 6274 16C 2.200GHz, Cray Gemini interconnect, NVIDIA 2090 Cray Inc.
7	CINECA Italy	**Fermi** - BlueGene/Q, Power BQC 16C 1.60GHz, Custom IBM
8	Forschungszentrum Juelich (FZJ) Germany	**JuQUEEN** - BlueGene/Q, Power BQC 16C 1.60GHz, Custom IBM
9	CEA/TGCC-GENCI France	**Curie thin nodes** - Bullx B510, Xeon E5-2680 8C 2.700GHz, Infiniband QDR Bull
10	National Supercomputing Centre in Shenzhen (NSCS) China	**Nebulae** - Dawning TC3600 Blade System, Xeon X5650 6C 2.66GHz, Infiniband QDR, NVIDIA 2050 Dawning

IT'S IN PROGRESS, AND THE CHANGES ARE MAJOR

A Brief History of (HP) Computing

Era	Mainframe	Vector	Parallel	Multi-level memory constrained
Time frame	60s to mid-70s	mid-70s to early-90s	Early 90s to early 10s	Early 10s to mid-20s
Fundamental scale	System in a room	System in a chassis	System on a board	System on a chip
Free performance	---	16x	40x	1x
Dominant constraint	Floating point capability	Physical size of system	Interconnect scalability	Energy efficiency, concurrency
Architectural challenges	---	Scatter-gather	Interconnect	Power Memory size Data motion Resiliency Heterogeneity
Programming model	Sequential processes	Vectorized sequential	Communicating sequential	Hierarchical parallel
Programming challenges	Expression of mathematical algorithms	Vectorized instructions & loops	Distributed applications & message passing	Data motion Multi-level parallelism Resiliency
Fundamental building block	Commercial CPUs	Custom CPUs	Commodity micro-processors	Heterogeneous cores
R&D initiatives	---	Seymour Cray, DOE procurements	HPCC, ASCI	Exascale
Country	U.S.	U.S., Japan	U.S.	China, EU, India, Japan, U.S.?

Courtesy of Sudip Dosanjh

Potential System Architecture Targets With Investment (2010 Projection)

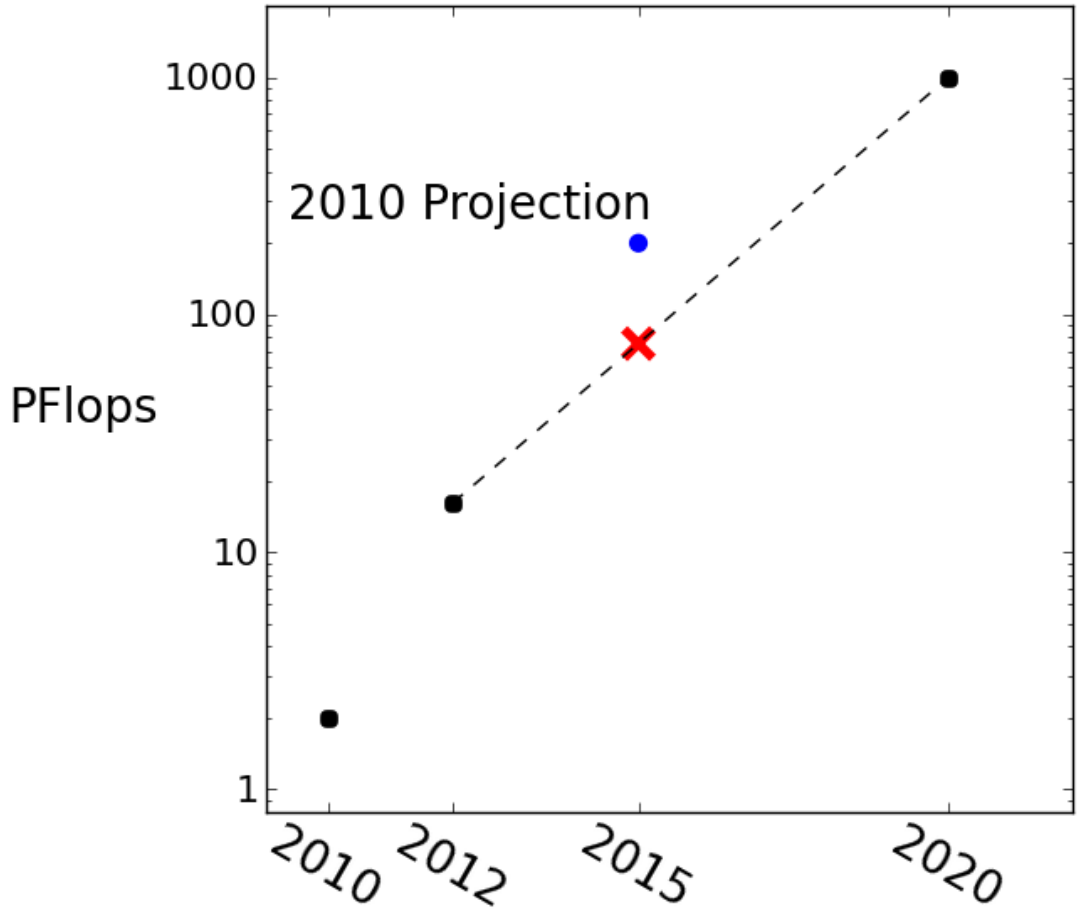
System attributes	2010	"2015"		"2018"	
System peak	2 Peta	200 Petaflop/sec		1 Exaflop/sec	
Power	6 MW	15 MW		20 MW	
System memory	0.3 PB	5 PB		32-64 PB	
Node performance	125 GF	0.5 TF	7 TF	1 TF	10 TF
Node memory BW	25 GB/s	0.1 TB/sec	1 TB/sec	0.4 TB/sec	4 TB/sec
Node concurrency	12	O(100)	O(1,000)	O(1,000)	O(10,000)
System size (nodes)	18,700	50,000	5,000	1,000,000	100,000
Total Node Interconnect BW	1.5 GB/s	20 GB/sec		200 GB/sec	
MTTI	days	O(1day)		O(1 day)	

Courtesy of Sudip Dosanjh

Exascale Changes Everything

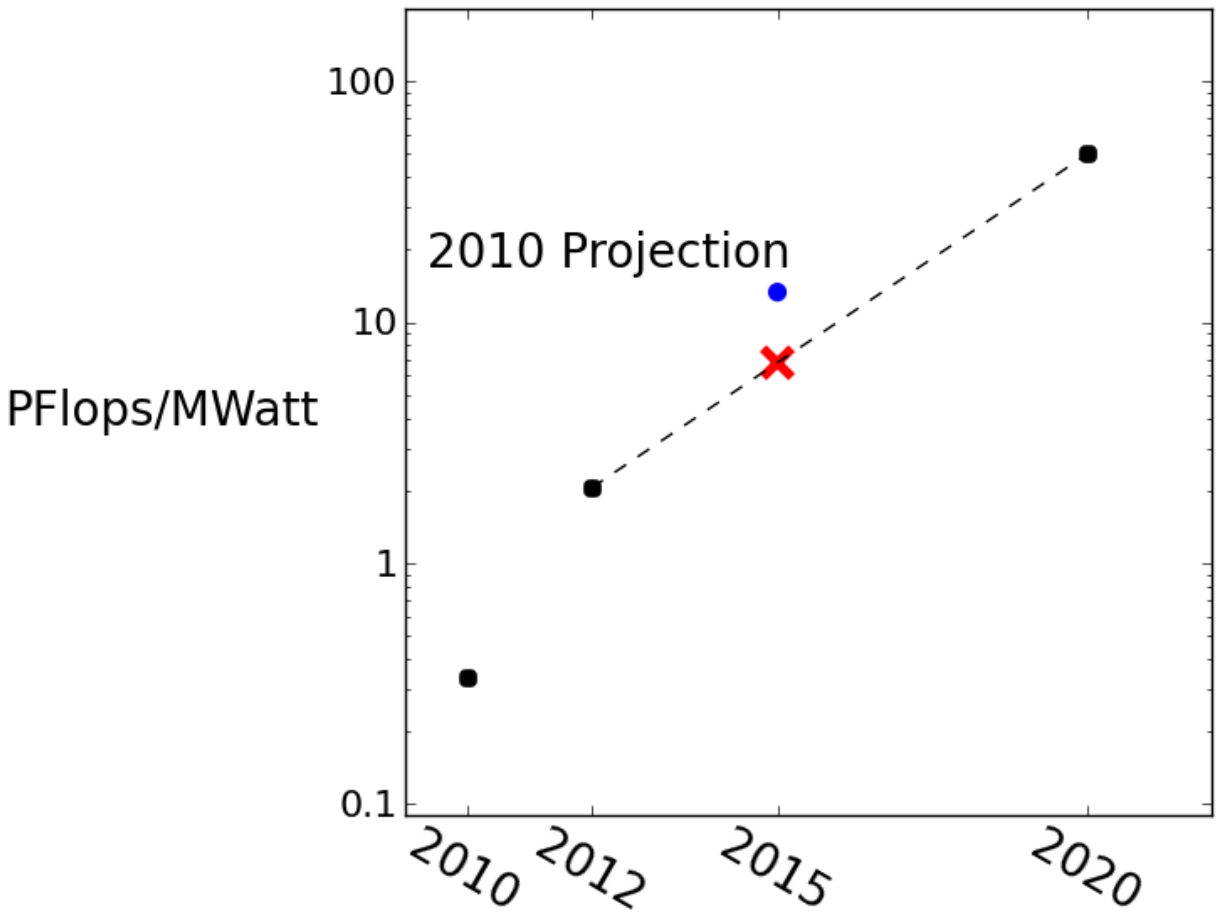
				Factor Change	
	2010	2012	2020	2010-2020	2012-2020
System Peak	2 Pf/s	16 Pf/s	1 Ef/s	500	62.5
Power	6 MW	7.8 MW	20 MW	3	2.56
Flops / Power	0.33 Gf/W	2 Gf/W	50 Gf/W	151	25
System Memory	0.3 PB	1.6 PB	10 PB	33	6.25
Node Concurrency	12 cpus	16 cpus	1,000 cpus	83	62.5
System Size (Nodes)	20 K nodes	98 K nodes	1 M nodes	50	10.2
Total Concurrency	225 K	1.5 M	1 B	4444	635
Storage	12 PB	50 PB	300 PB	20	6
Input/Output BW	0.2 TB/s	1.0 TB/s	20 TB/s	100	20

Flop rate is on track for 2020



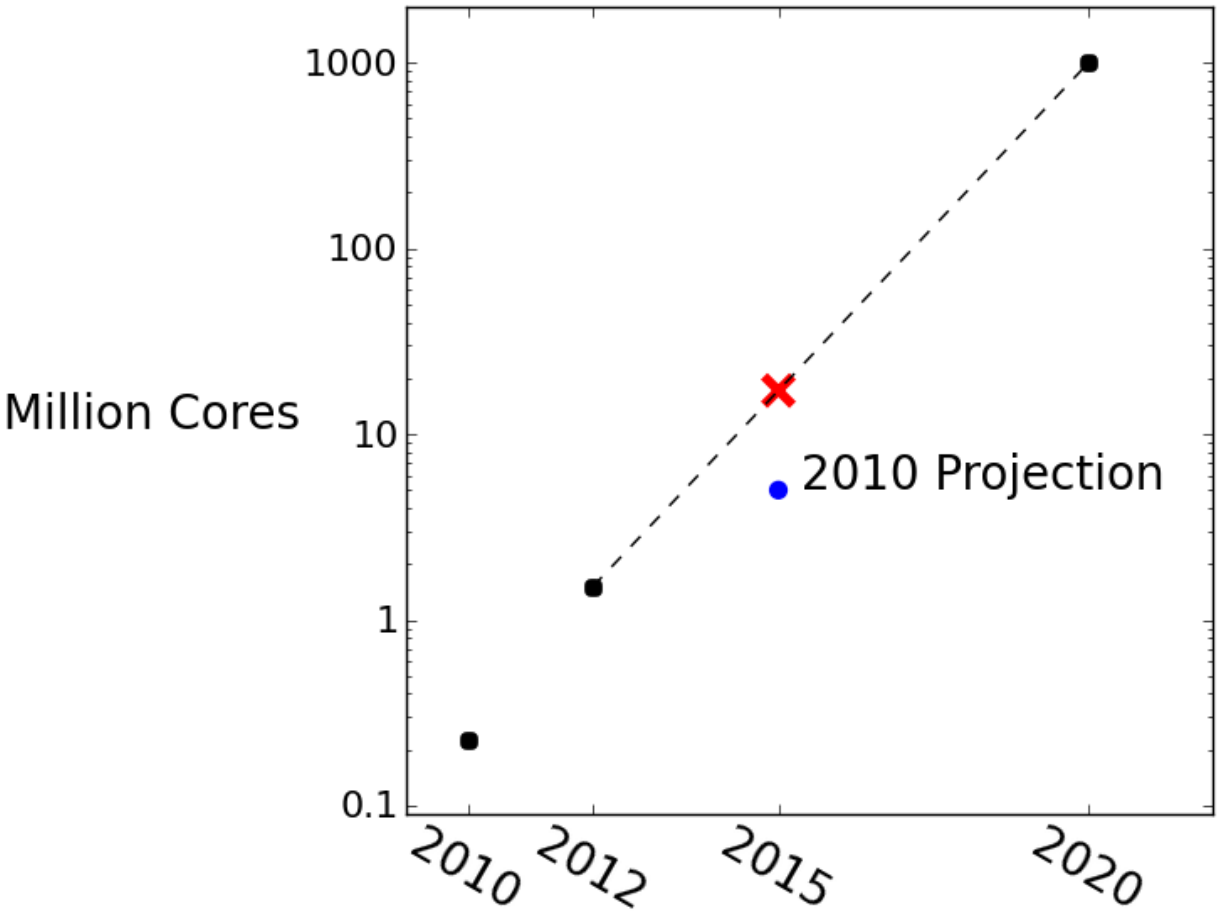
Sandia National Laboratories

Flop/MW ratio is on track for 2020



Sandia National Laboratories

Concurrency is on track for 2020



Sandia National Laboratories

Heterogeneous multicore nodes are our future

Home | Product Guides | CPUs & Components | Chipsets & Processors | NVIDIA To Break Into PC CPU Market

NVIDIA To Break Into PC CPU Market

Project Denver CPU is an ARM-based chip for "high-performance" computing.

By Sascha Segan | January 5, 2011 04:54pm EST | 1 Comment | Email | Print

+1 1 | Share 10 | Tweet 11 | Share | 1 | Digg | Submit

more CES2011

LAS VEGAS -- NVIDIA is working on "Project Denver," a high-performance ARM-based processor designed for PCs, servers, and supercomputers, CEO Jen-Hsun Huang said at the CES trade show today.

"What we're building is a full, custom processor developed at NVIDIA in partnership with ARM, that is based on ARM," Huang said. "This is the world's first ARM processor targeted at high-performance computing."

ScaleMP sgi SIMULIA

Since 1986 - Covering the Fastest Computers in the World and the People Who Run Them

Visit additional Tabor Communication Publications

Home | News | Features | Blogs | HPC Markets | Whitepapers | Multimedia

Digital Manufacturing report

HPC In the Cloud

2011 HPCwire Readers Choice Awards

August 15, 2011

Adventures with HPC Accelerators: GPUs and Intel MIC Coprocessors

Aaron Dubrow, Texas Advanced Computing Center

Researchers from Mellanox Technologies and the Texas Advanced Computing Center share early experiences at TeraGrid '11

For the past few years, the buzz around hardware accelerators, particularly graphics processing units (GPUs), has been growing. Designed with a massive number of floating point units and very high memory bandwidth so as to accelerate certain computing processes, GPUs and other emerging accelerates

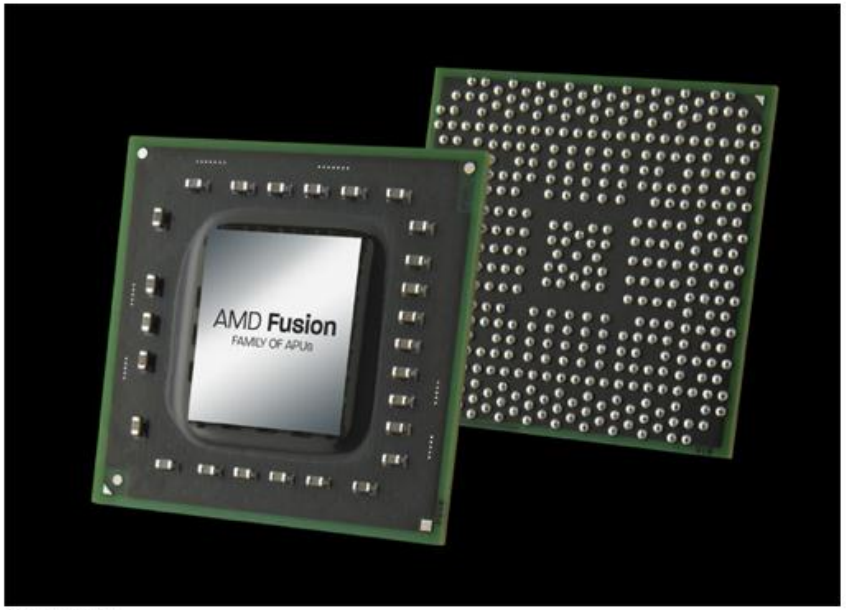
AMD Fusion laptop roundup

By: Dan Ackerman

MARCH 24, 2011 11:50 AM PDT

Print E-mail

Recommend 25 | Tweet 42 | +1 2 | Share | 6 comments

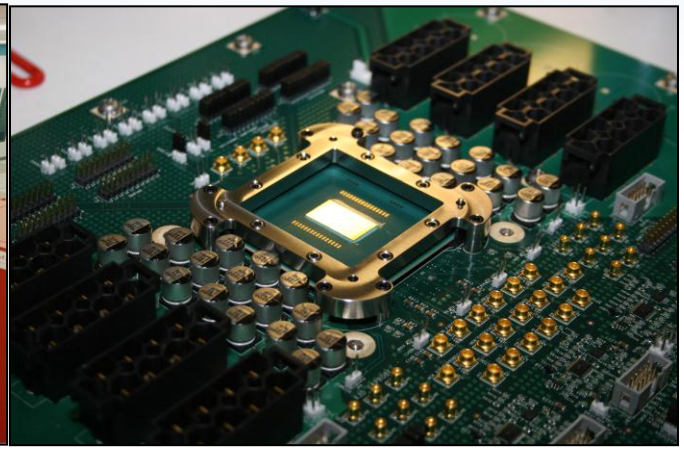


(Credit: AMD)

Courtesy of Sudip Dosanjh

Future microprocessors will be much more difficult to use than ASCI Red

- 1.8 TF
- Homogenous
  - 4500 nodes
- DRAM
  - Capacity – 0.6 TB
  - Bandwidth – 2.4 TB/s
- 2500 sq ft
- 0.5 MW



- 10 TF
- Multilevel & Heterogeneous
  - ~10 CPU cores,
  - ~1000 GPU cores
    - 100 threads/core
- DRAM
  - 0.3 TB
  - 0.1 TB/s
- Size of a dime
- 200 W

10X lower byte/FLOP and 100X lower memory BW/FLOP

In addition, we need 100,000 10 TF processors to build an Exascale system

Courtesy of Sudip Dosanjh

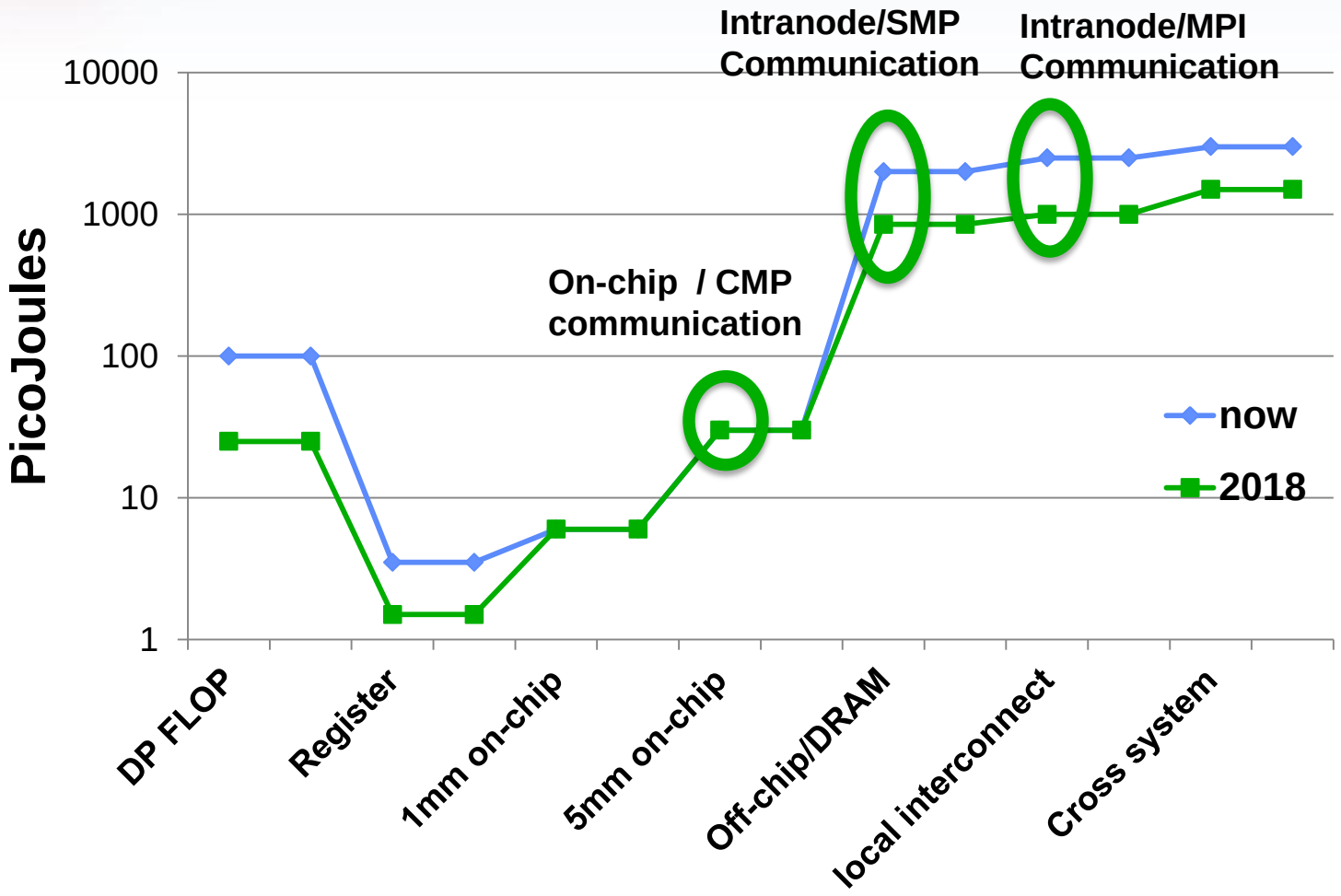
CHALLENGES

… but it won’t be easy.

Key Technical Challenges

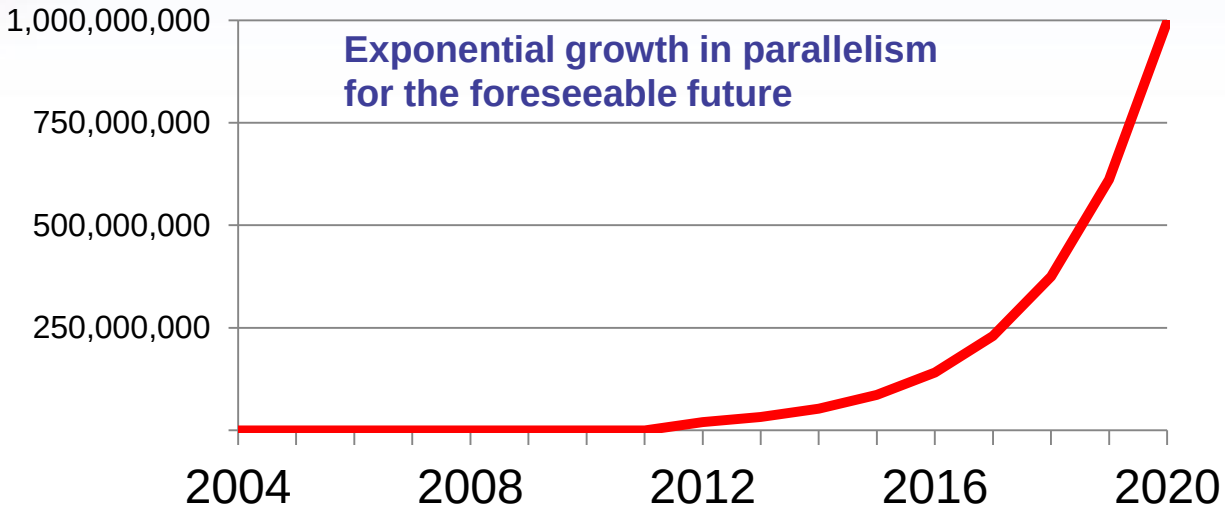
- **DOE's Exascale Initiative Steering Committee and DARPA identified technology gaps that need to be addressed to reach Exascale later this decade**
  - **Power, memory and storage, parallelism and locality, resilience, scalability, programming models**
- **Co-development (or co-design) of hardware, system software and applications is a key element of our strategy**
  - **Codes will need to adapt to manage billion-way parallelism, data locality, resilience and perhaps energy**

Investments in architecture R&D and application locality are critical



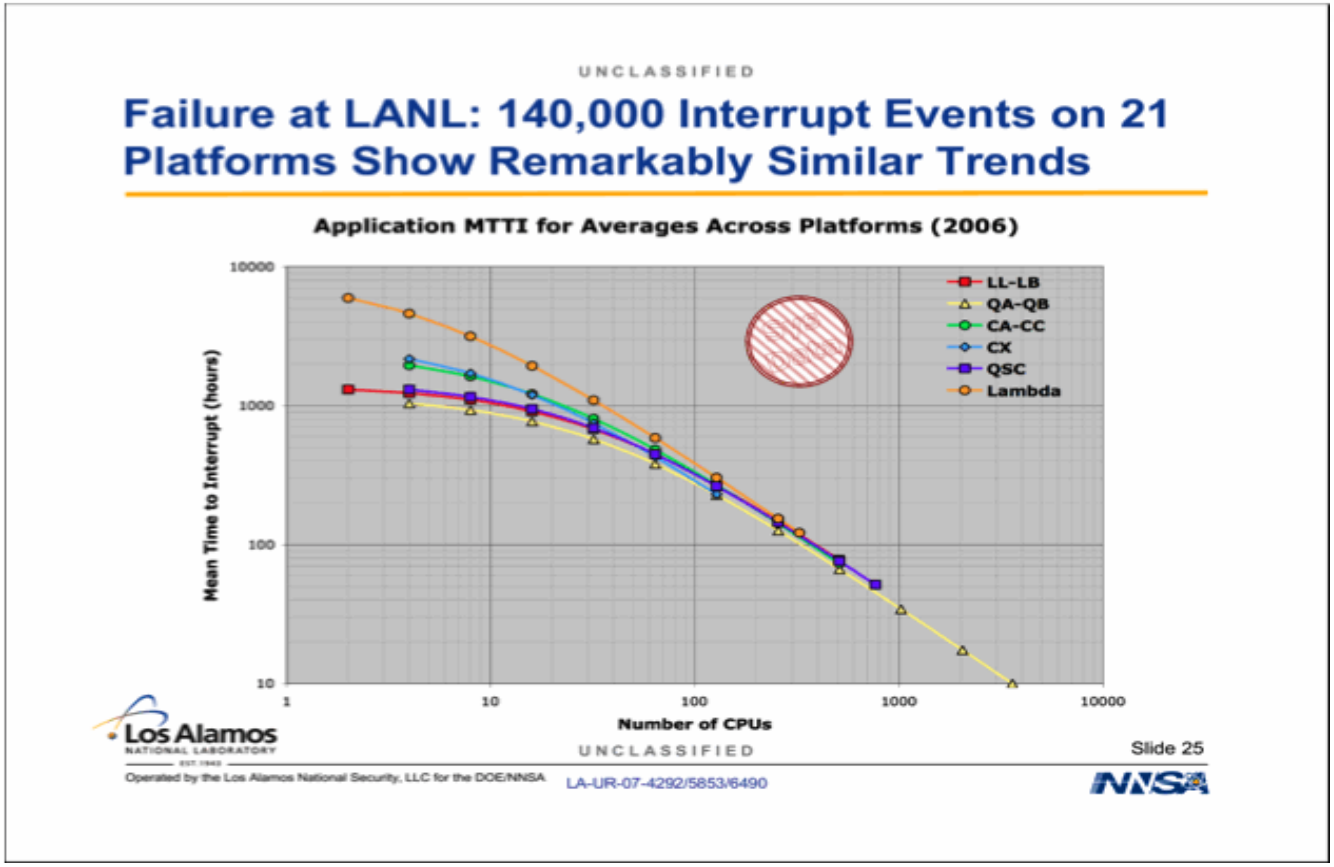
"The Energy and Power Challenge is the most pervasive … and has its roots in the inability of the [study] group to project any combination of currently mature technologies that will deliver sufficiently powerful systems in any class at the desired levels."
DARPA IPTO exascale technology challenge report

Parallelism is Exploding



- **Billion-way parallelism at Exascale**
  - **Many levels of parallelism**
  - **Heterogeneity**
- **Fundamentally breaks scaling assumptions of current software**
- **Drives the need for a new programming models**
- **Weak scaling will not be sufficient**
- **Impacts the entire HPC ecosystem**

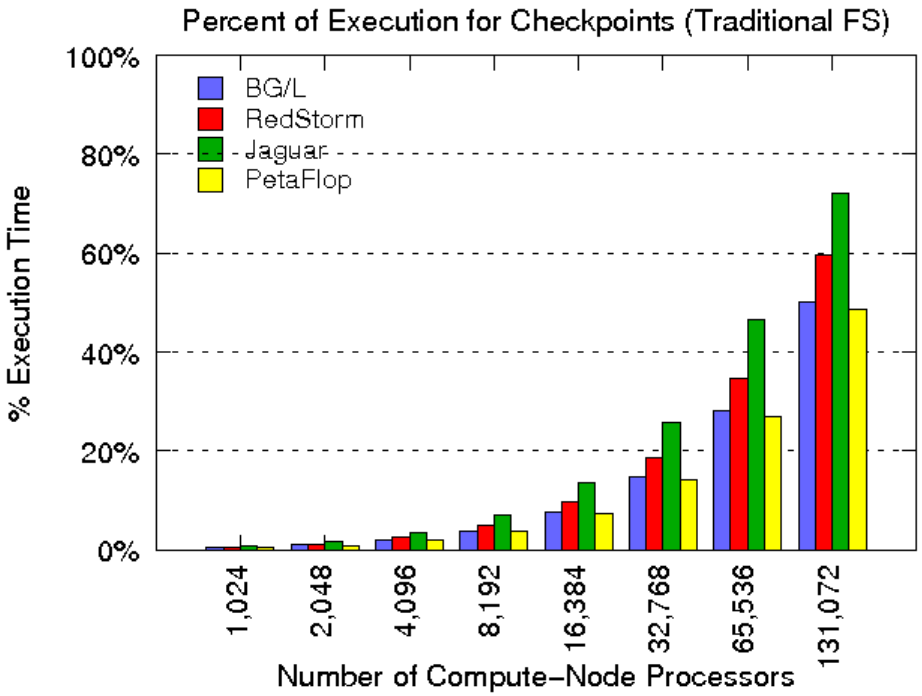
MTTI is shrinking as # cores grows



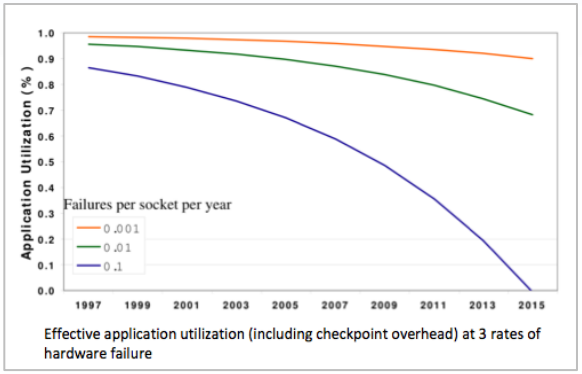
(Courtesy of John Daly)

Sandia National Laboratories

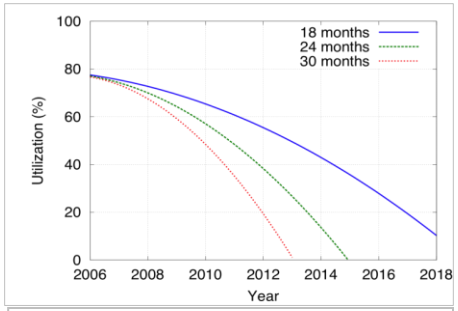
Checkpoint trend isn't good



Oldfield et al., *Modeling the Impact of Checkpoints on Next-Generation Systems*. MSST, 2007



(Courtesy of Lucy Nowell & Sonia Sachs)



Schroeder and Gibson, *Understanding Failures in Petascale Computers*. Journal of Physics, 2007

(assuming that the number of cores per socket grows by a factor of 2 every 18, 24 and 30 months)

Machine utilization is going to zero! (Not really)

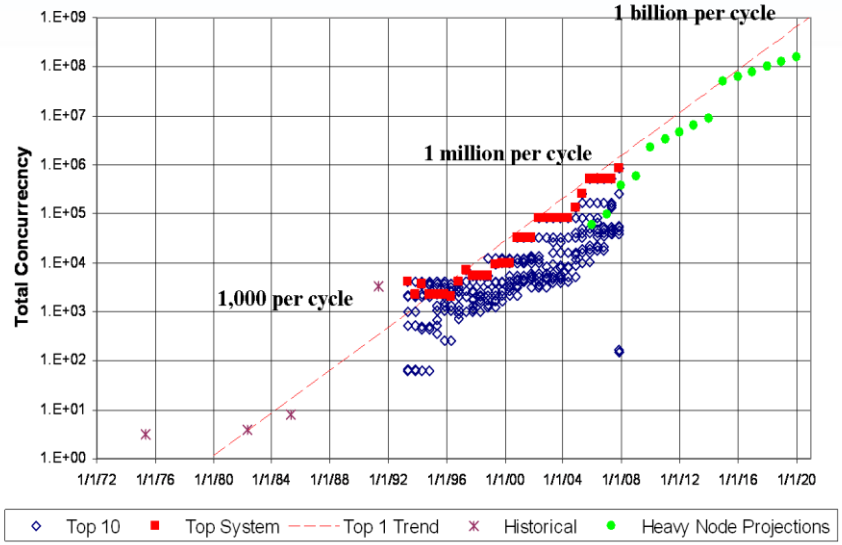
Sandia National Laboratories

Programming Models

- What is required of the programming model?
- Why can't we just use the same one we're using today?
- Is there a clear winner? Will there be 'just one'? [composition]
- What's changing?
  - CPUs
  - GP-GPUs?
  - FPGAs?
  - Custom accelerators?
- Hierarchical memory layers growing
  - On chip (shared cache?)
  - On board (shared memory?)
  - Inter-node
  - SSD (on-node)
  - HD
- What models are out there?
- Cite LLNL study

Programming Models and Environments

- **Barriers:**
  - **O(1B) way parallelism in Exascale system**
    - **Massive lightweight cores for low power**
    - **Some "full-feature" cores lead to heterogeneity**
  - **O(10K) way parallelism in a processor**
    - **Data and independent thread parallelism**
  - **Data movement costs power and time**
    - **Software-managed memory (local store)**
  - **Programming for resilience**
  - **Science goals require complex codes**
- **Technology Investments**



How much parallelism must be handled by the program?
From Peter Kogge (on behalf of Exascale Working Group), "Architectural *Challenges* at the Exascale Frontier", June 20, 2008

- **Extend existing between-chip models for scalability and resilience, e.g., MPI with support to hide hardware failures and low memory footprint**
- **Develop on-chip models for 10K-way concurrency and heterogeneity by adapting current ones (e.g., OpenMP) or leverage models from other domains (e.g., CUDA or OpenCL)**
- **Revolutionary: enable new software model for high concurrency across system scales**

- **Technical Gap: Productivity, performance and correctness for 1000x more parallelism on chip while increasing programming productivity of scientists by 10x**

System complexity is increasing

- **Heterogeneous node architecture (accelerators)**
  - **CPUs**
  - **GP-GPUs?**
  - **FPGAs?**
  - **Custom accelerators?**
- **Hierarchical memory layers growing**
  - **On chip (shared cache?)**
  - **On board (shared memory?)**
  - **Inter-node**
  - **SSD (on-node)**
  - **HD**

Apps developers have some serious challenges ahead

- **Extreme parallelism – concurrency is everything**
- **High cost of moving data – locality is everything**
- **Reduced system (HW & SW) reliability – static model is broken; need to think dynamic model**
- **“Unknown” architectures – variations on themes**
- **Programming model is unclear**

Code rewrite could be a huge deal, and take years. It would be good to understand how codes would run on various architectures.

SELECTED WORK AT SANDIA

Fault-oblivious computing, performance simulation, and scalable data analysis.

Work of Jackson Mayo, David Thompson, Janine Bennett, Rob Armstrong

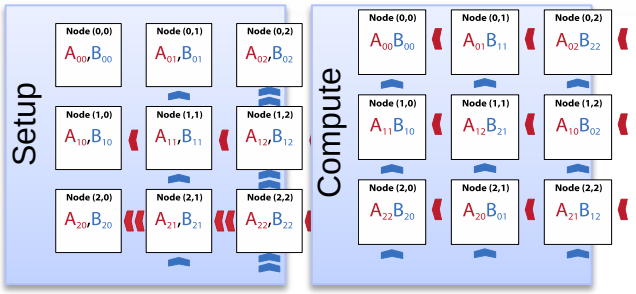
FAULT-TOLERANCE IS AN APPLICATION ISSUE

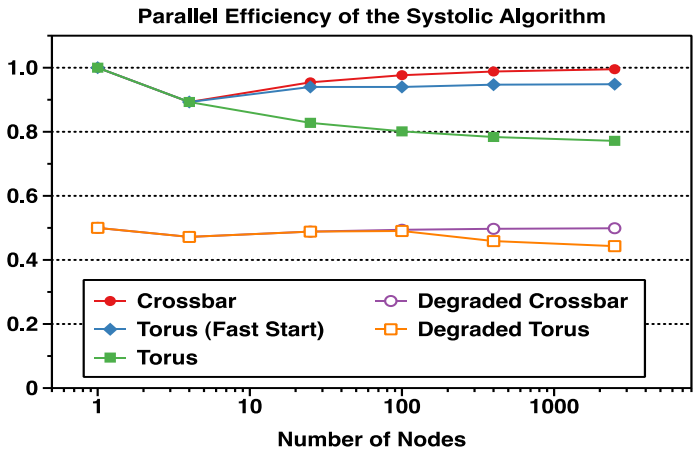
Existing SPMD programming models are inherently NOT fault tolerant

The move to exascale only makes things worse

- Global checkpoints no longer feasible
- Global collectives costly
- Applications/runtime must handle soft and hard failures
- Asynchronous execution to hide memory & I/O latency
- Deep memory hierarchies require tuning

Example: Systolic Matrix Multiplication





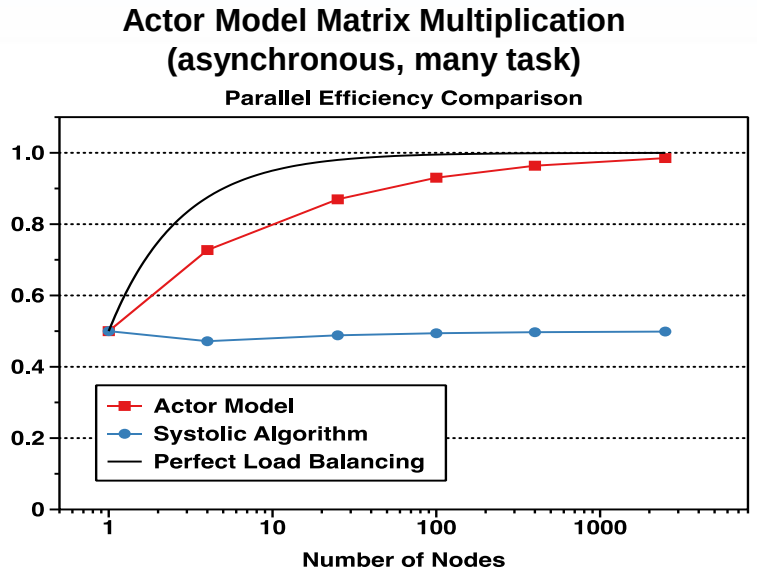
The implicitly synchronous systolic algorithm cannot recover from node degradation

C. L. Janssen, H. Adalsteinsson, J. P. Kenny, *Using simulation to design extreme-scale applications and architectures: programming model exploration*, ACM SIGMETRICS Performance Evaluation Review, **38**, pp. 4-8, 2011.

Sandia National Laboratories

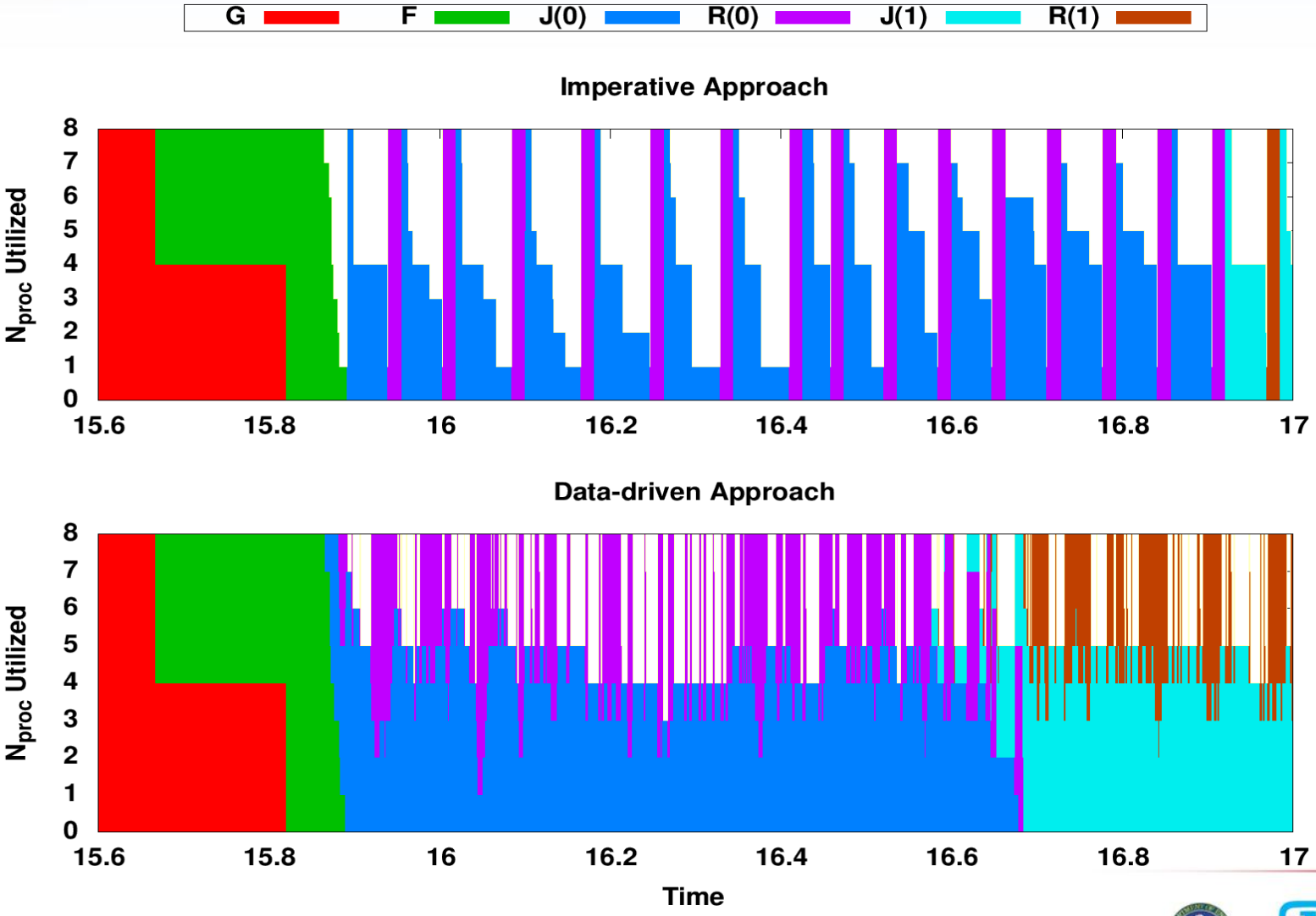
Asynchronous many-task programming models are fault tolerant!

- Simulation permits straightforward investigation of alternative programming models
- Work-stealing approaches will play a role in dealing with large-scale machines lacking perfect homogeneity



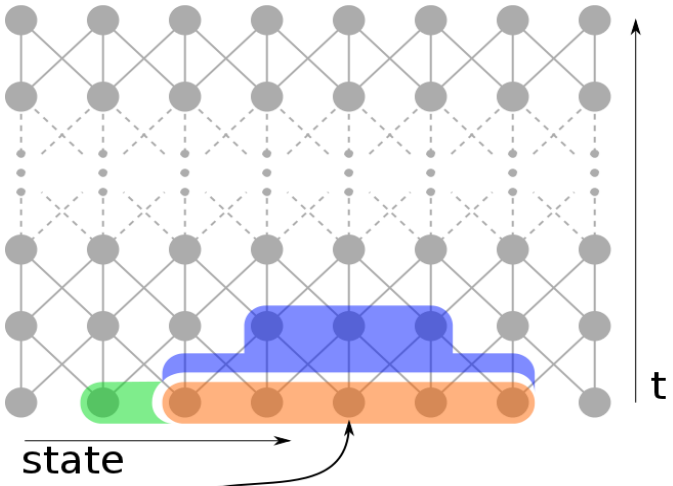
- Research Questions:
  - Is MPI+X (*global* checkpoint/restart) enough?
  - If not, what programming models can reach what scales?
  - If no programming model can reach scales of interest for a given application without algorithmic changes, how might algorithms be adapted?
  - Codesign of architecture tradeoffs between memory, I/O, power, and application performance

Simulated timings for 16 shells on 8 processors

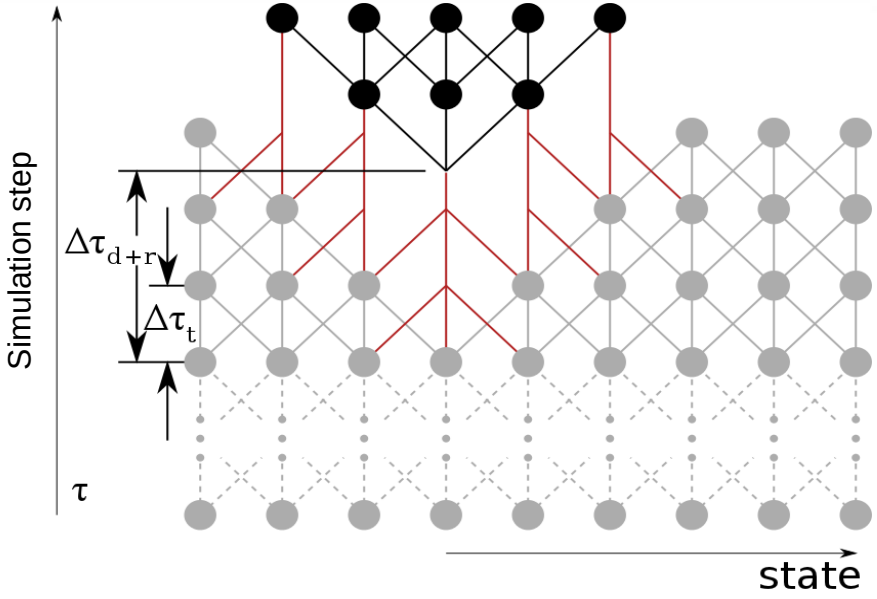


Sandia National Laboratories

Fritz: A task-DAG programming model for scaling amid failures



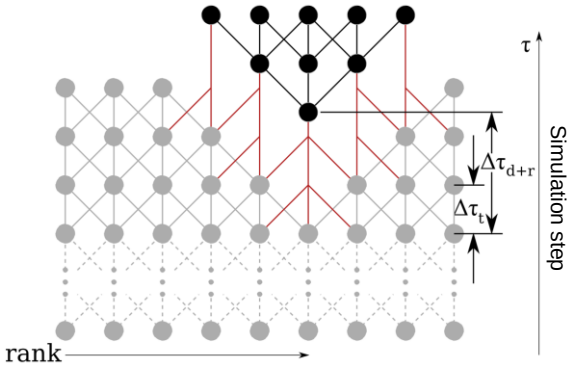
Each process holds a subset of state. Process holdings overlap entirely for redundant, in-memory recovery.

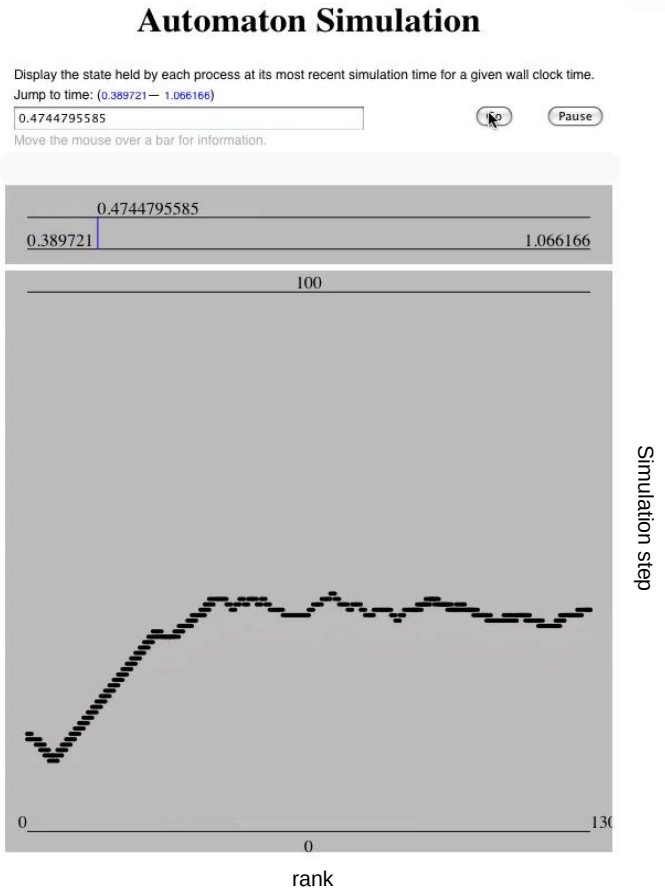


Failures & delays propagate from process to process as state dependencies are communicated.

Task-DAGs provide resilience, but challenges remain

- Development of resilient decentralized scheduling algorithms
- Robust generation of missing subgraph upon failure with local (disk) checkpointing
- Intuitive expression of tasks to achieve asynchronous execution
- Work-sharing to avoid additive cascading delays

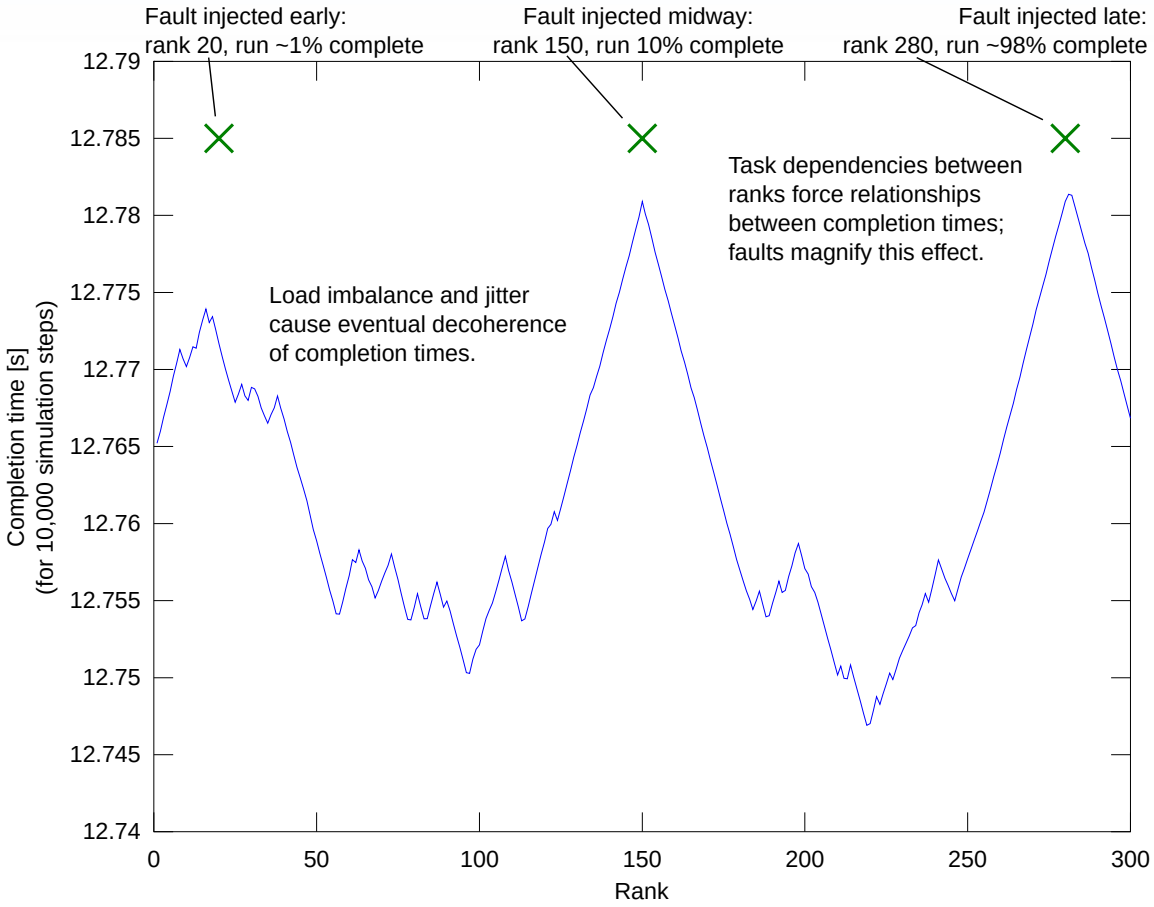




The movie shows a sequence of horizontal slices through the diagram on the left. Each slice indicates to which simulation step each process has advanced.

Sandia National Laboratories

Delay scaling of Fritz shows promise under a Poisson failure model

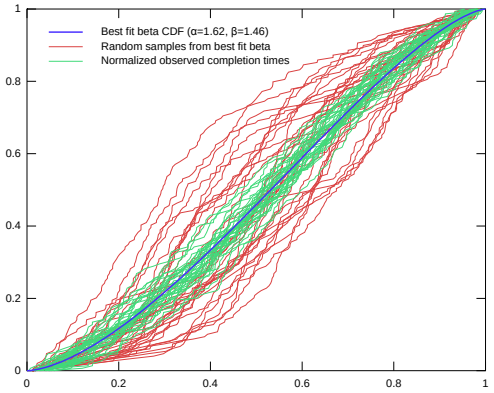


Snapshot of three failure + recovery delays induced on different processes at different simulation time steps

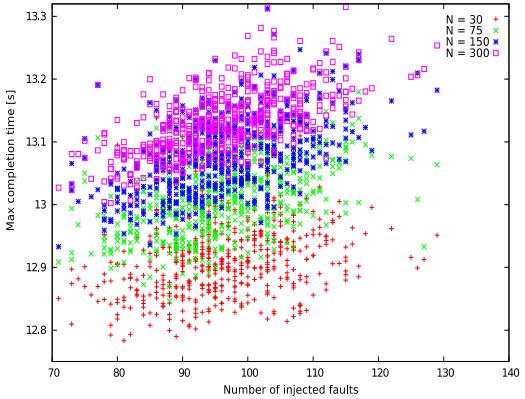
Sandia National Laboratories

Initial studies show that average maximum completion time is growing sub-logarithmically

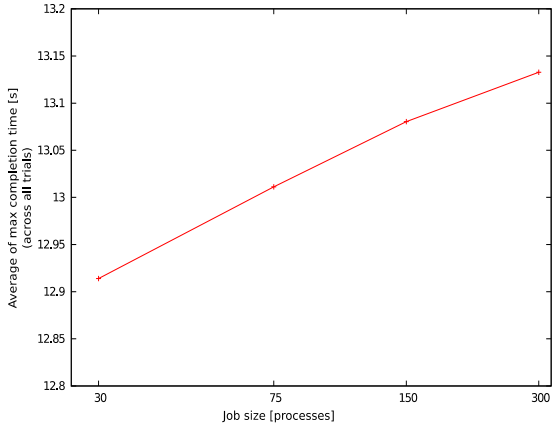
CDFs of completion times



Max completion vs. # faults



Average max completion vs. job size

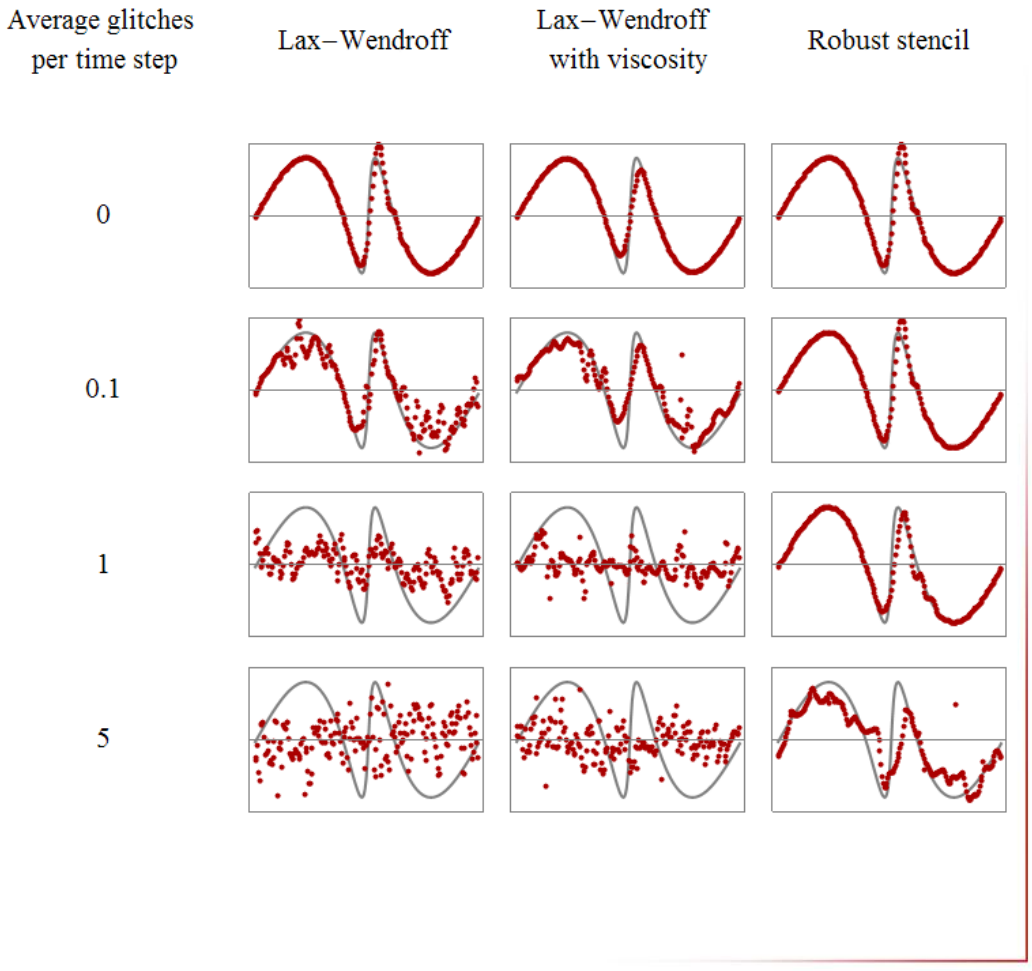


- This is good news for local recovery strategies
  - it hints that delays do not compound each other as scale increases
- Larger scaling studies are in progress

Sandia National Laboratories

“Robust stencils” can discard outliers to mitigate bit flips in PDE solving

- A simple 1D advection equation $\partial u/\partial t = \partial u/\partial x$ illustrates the behavior of finite-difference schemes
- The robust stencil here computes a second-order update at position i from one of these subsets after discarding the most extreme value:
  - $\{ i-3, \quad i-1, \quad i+1, \quad i+3 \}$
  - $\{ \quad i-2, \quad i, \quad i+2 \quad \}$
  - $\{ \quad i-1, \; i, \; i+1 \quad \}$



Sandia National Laboratories

Work of Gilbert Hendry, Arun Rodriguez,Joe Kenny, Jeremiah Wilke, and Damion Dechev

ANALYZING THE PERFORMANCE BEFORE THE MACHINE EXISTS

Simulation in the Exascale Design Space

SST/macro

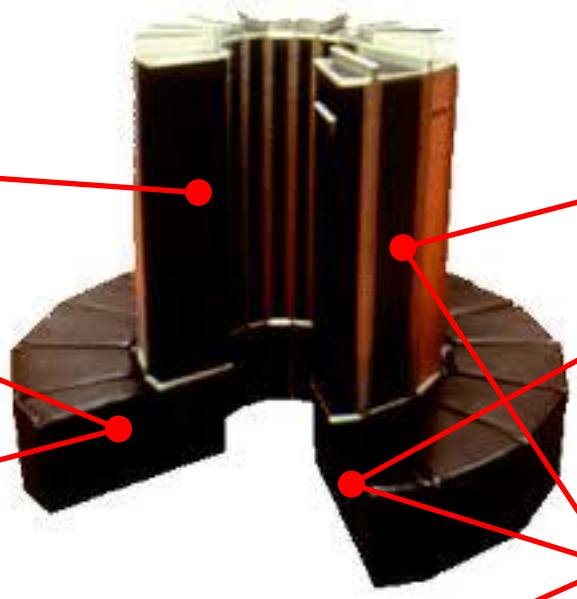
Network architecture/topology

Application development

Library/interface/services support

Machine fault tolerance

File system, I/O



(supercomputer)

SST/micro

Network switch implementation

Node Architecture

Memory technology

Processor/GPU

Power

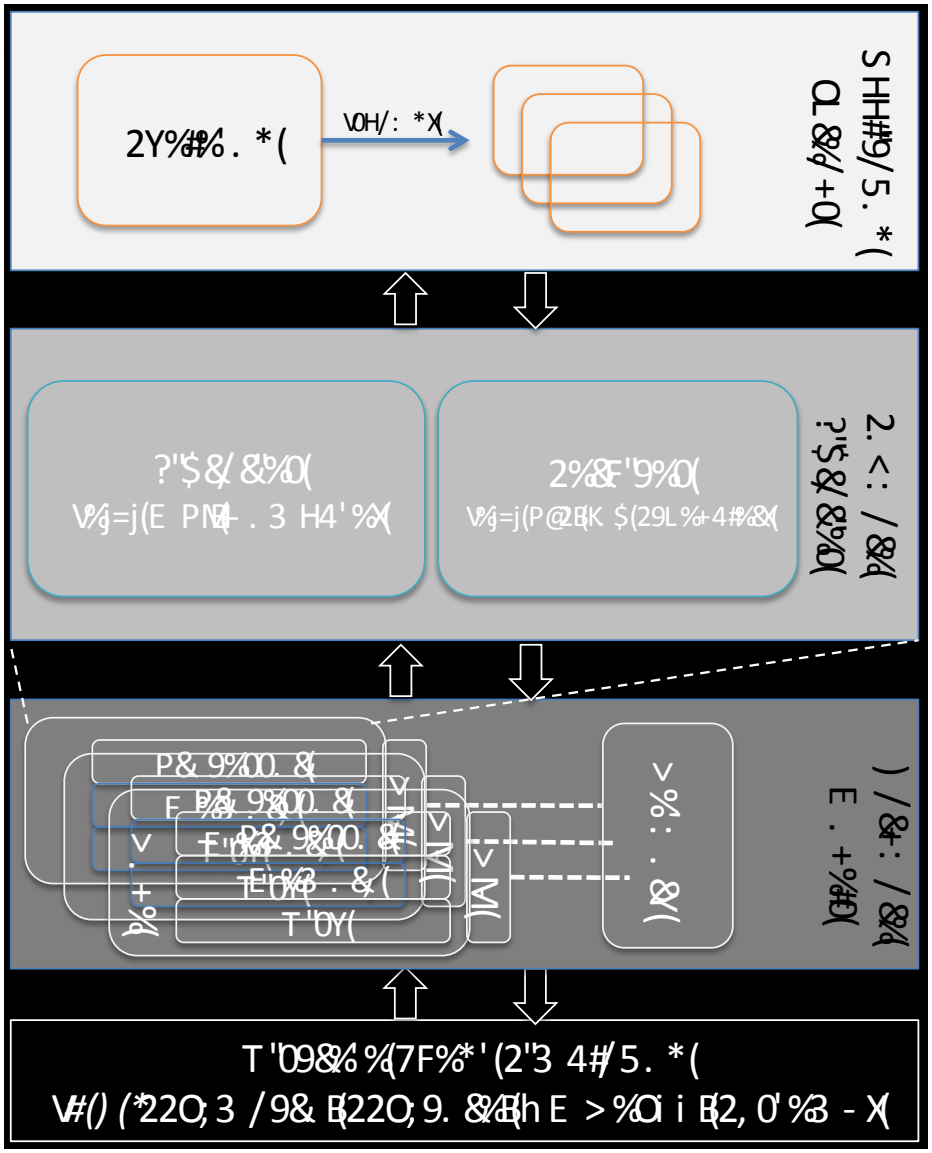
Meso-Scale Simulation

Ask implementation-related questions at scale.

SST/macro: Coarse-Grained Simulation

An application code with minor modifications

Our implementation of interfaces (MPI), which simulate execution and communication



Related Work:

xSim (Oakridge) – only MPI
BigSim (UIUC) – only MPI, needs a big machine

Mini Apps: An efficient vehicle for Co-Design

Mini-apps:

(sometimes called compact apps, or reduced app)

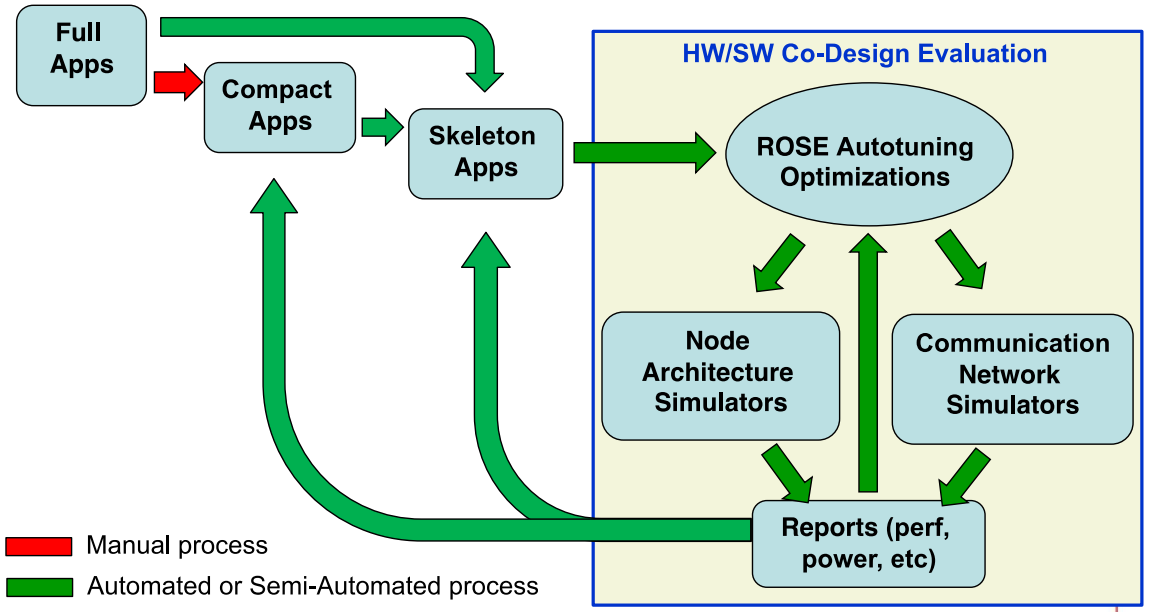
- A more efficient vehicle for co-design

Skeletons:

- A type of mini-app which strips out everything except communication and control
- Run 2-10x *faster* than real execution

(as opposed to 10-1000 *slower* in cycle-accurate)

- Design vehicle for fast prototyping



code:

```
int x = rank % 2;

int count = 100;
int* buf = (int*)malloc(count * sizeof(int));

if(x == 0){
    for(int i = 0; i < count; i++){
        buf[i] = rand();
    }
    MPI_Send(buf, count, MPI_INT, rank + 1, 0, MPI_COMM_WORLD);
}else{
    MPI_Recv(buf, count, MPI_INT, rank - 1, 0, MPI_COMM_WORLD);
}
```

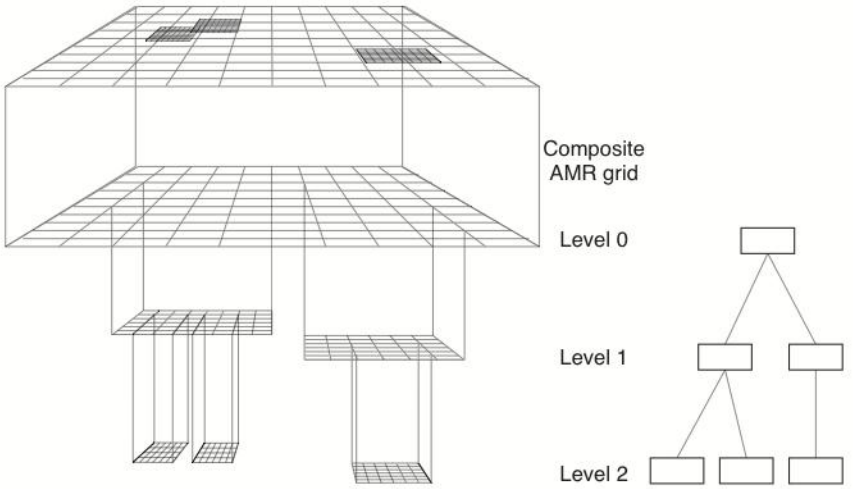
skeleton:

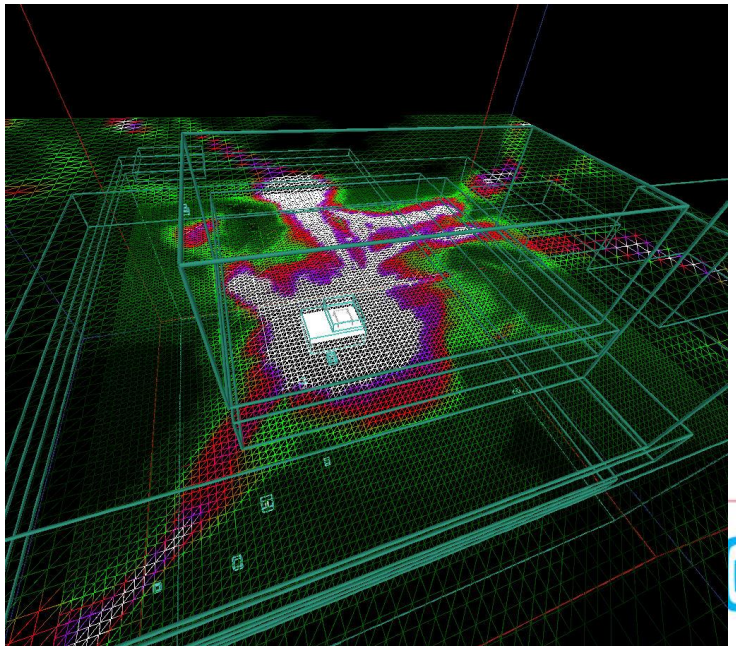
```
int x = rank % 2;

int count = 100;
/*model memory allocation here
 */
if(x == 0){
    /*model computation here
     */
    MPI_Send(NULL, count, MPI_INT, rank + 1, 0, MPI_COMM_WORLD);
}else{
    MPI_Recv(NULL, count, MPI_INT, rank - 1, 0, MPI_COMM_WORLD);
}
```

Robe

Using SST/macro - Combustion

- **Adaptive Mesh Refinement (AMR) important for efficiently looking at regions of space**
- **Very data-dependent, so hard to make a skeleton**
- **Porting parts of BoxLib (LBL) into SST/macro to investigate box generation and layout at large scales**

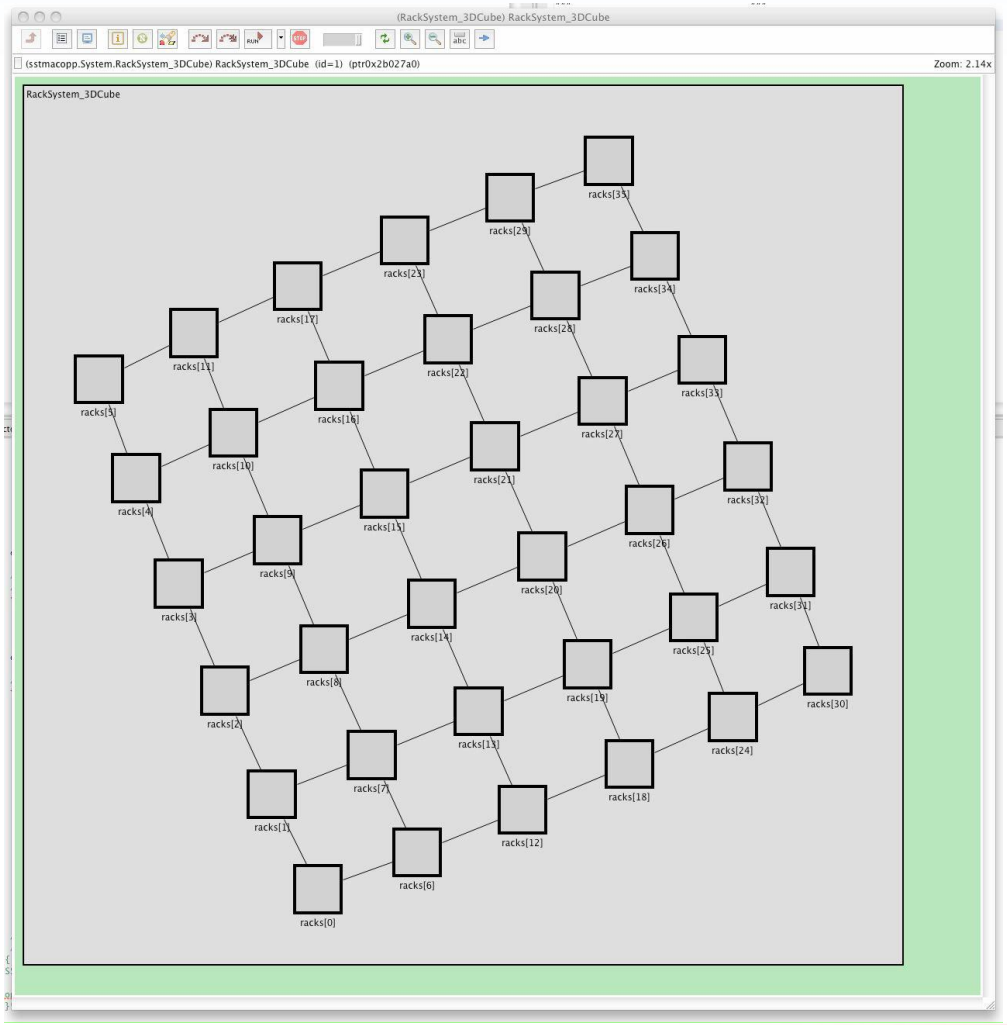




Experiment: Actor Load Balancing

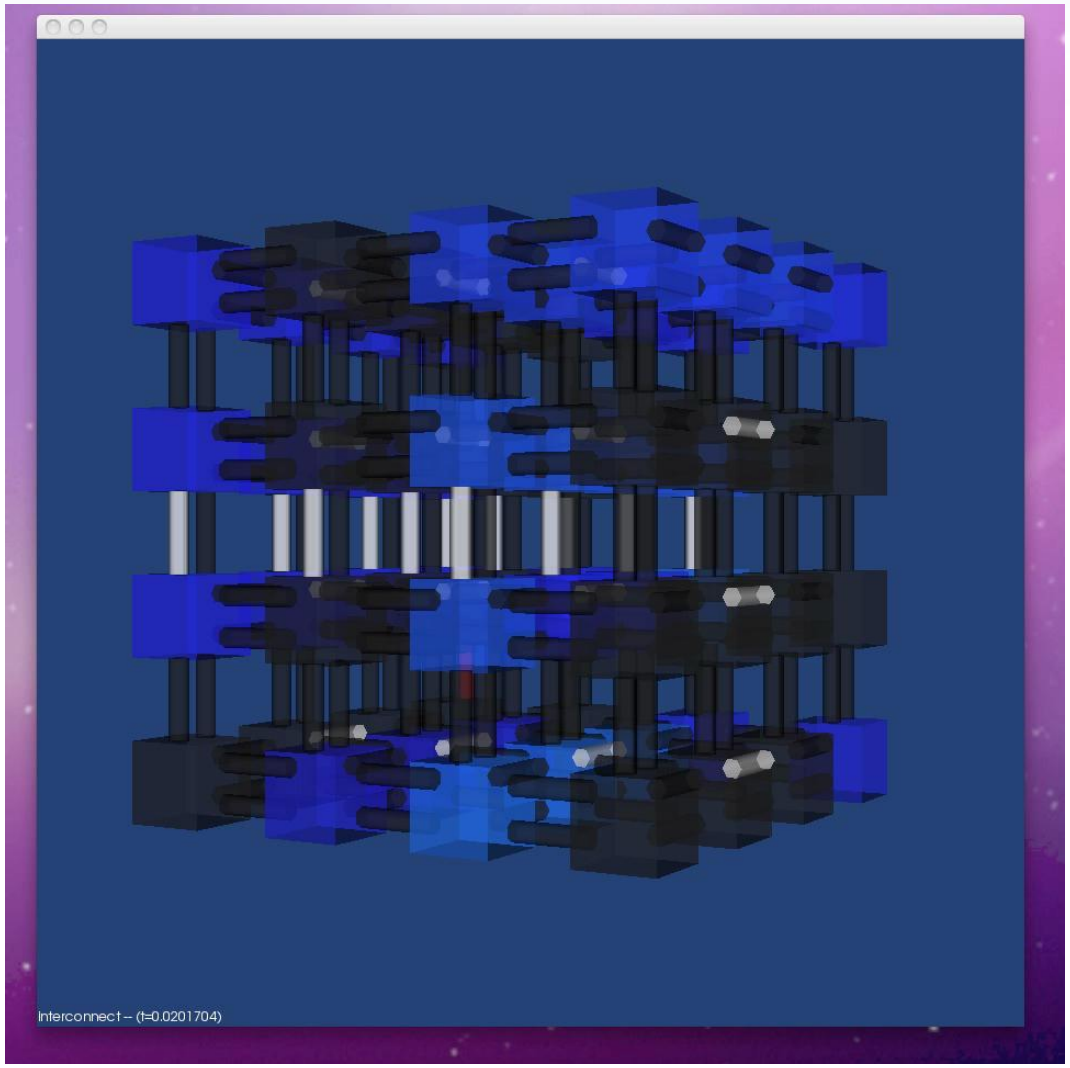
Legend

- Green – working
- Yellow border – prefetching
- Red – idle
- Purple – work stealing



Sandia National Laboratories

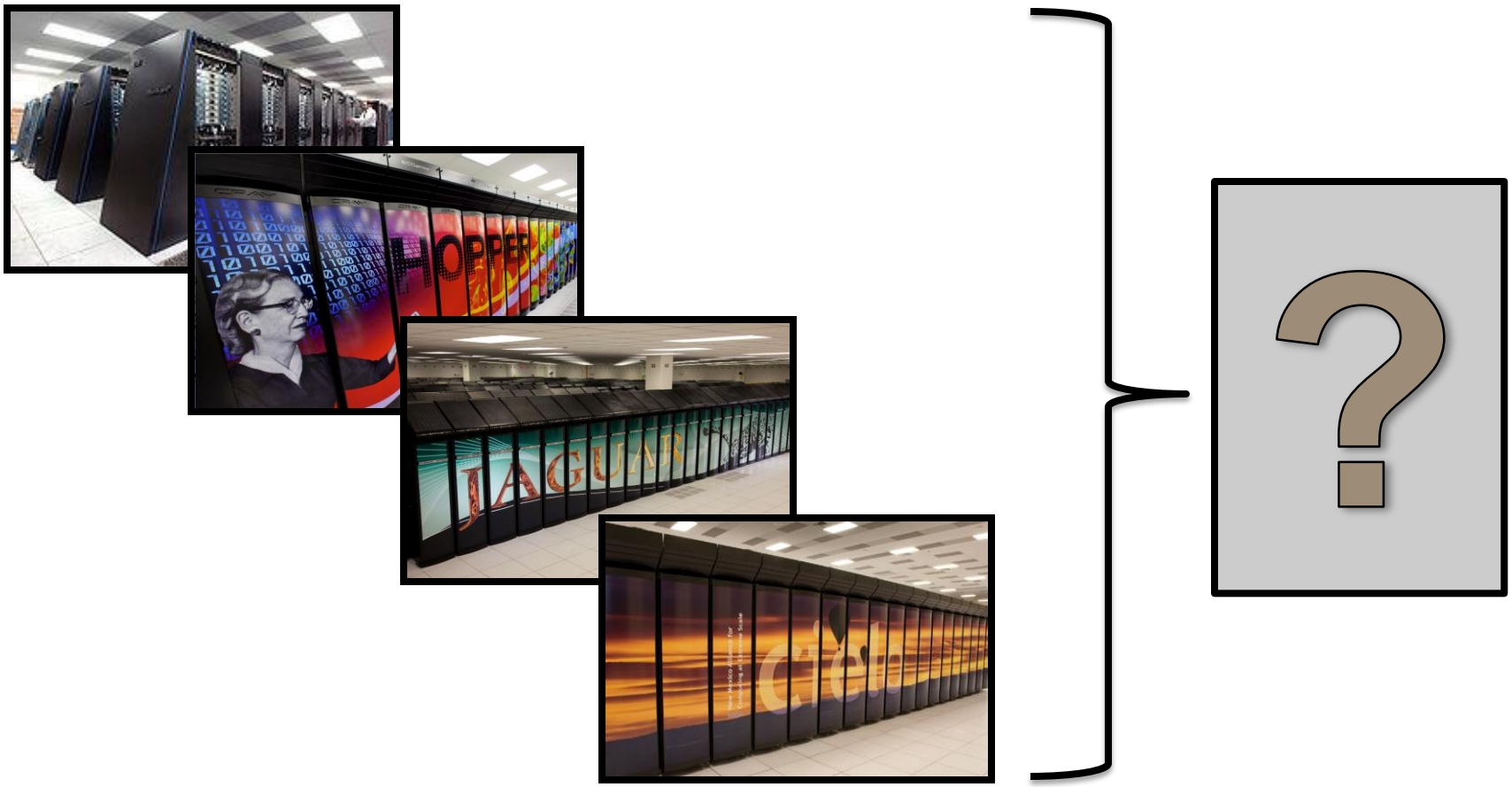
SST Network Traffic Visualization with VTK



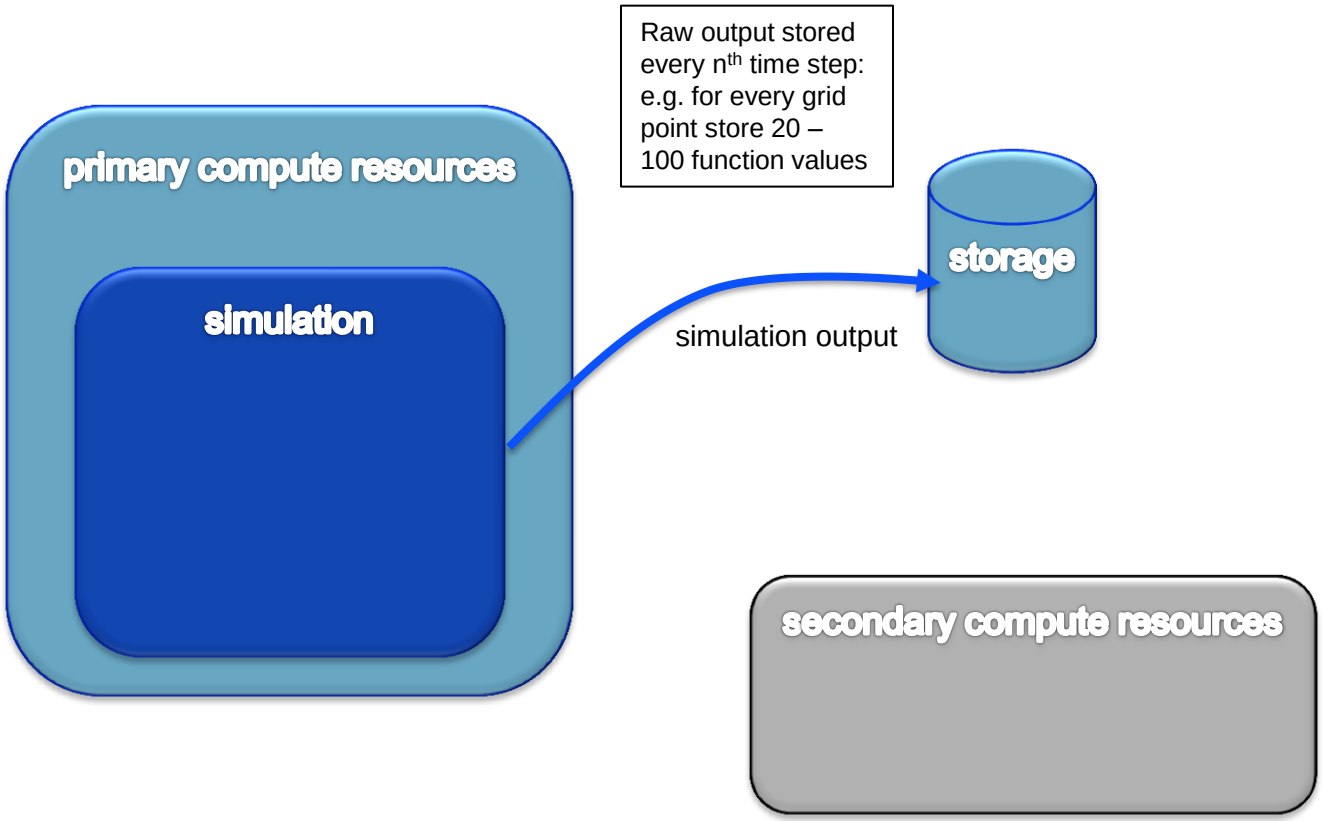
Work of Janine Bennett, Jackie Chen, Hongfeng Yu, Valerio Pascucci, Manish Parashar, …

DATA/RESULTS ANALYSIS IS CHANGING AS WELL

Extreme-scale data analysis on future architectures

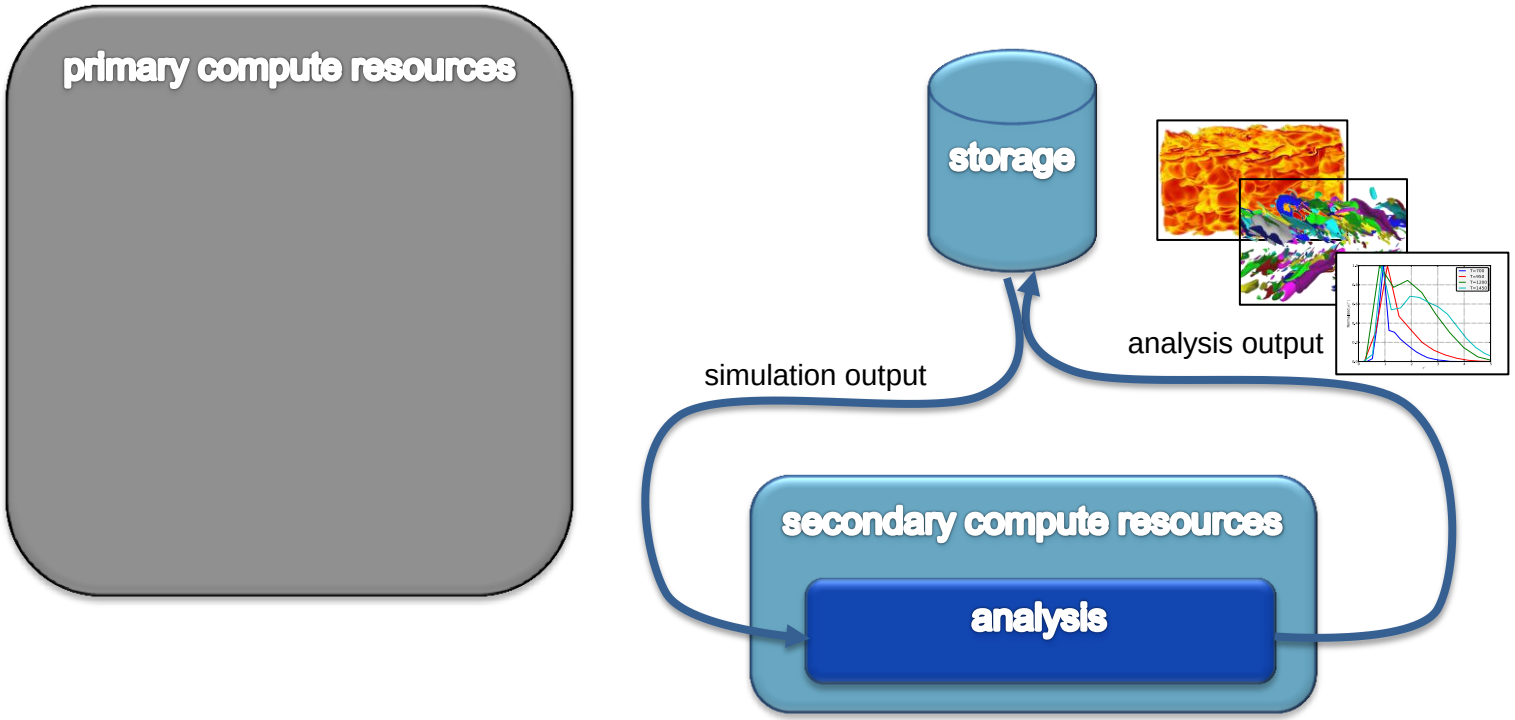


Existing data analysis paradigm comprises two stages



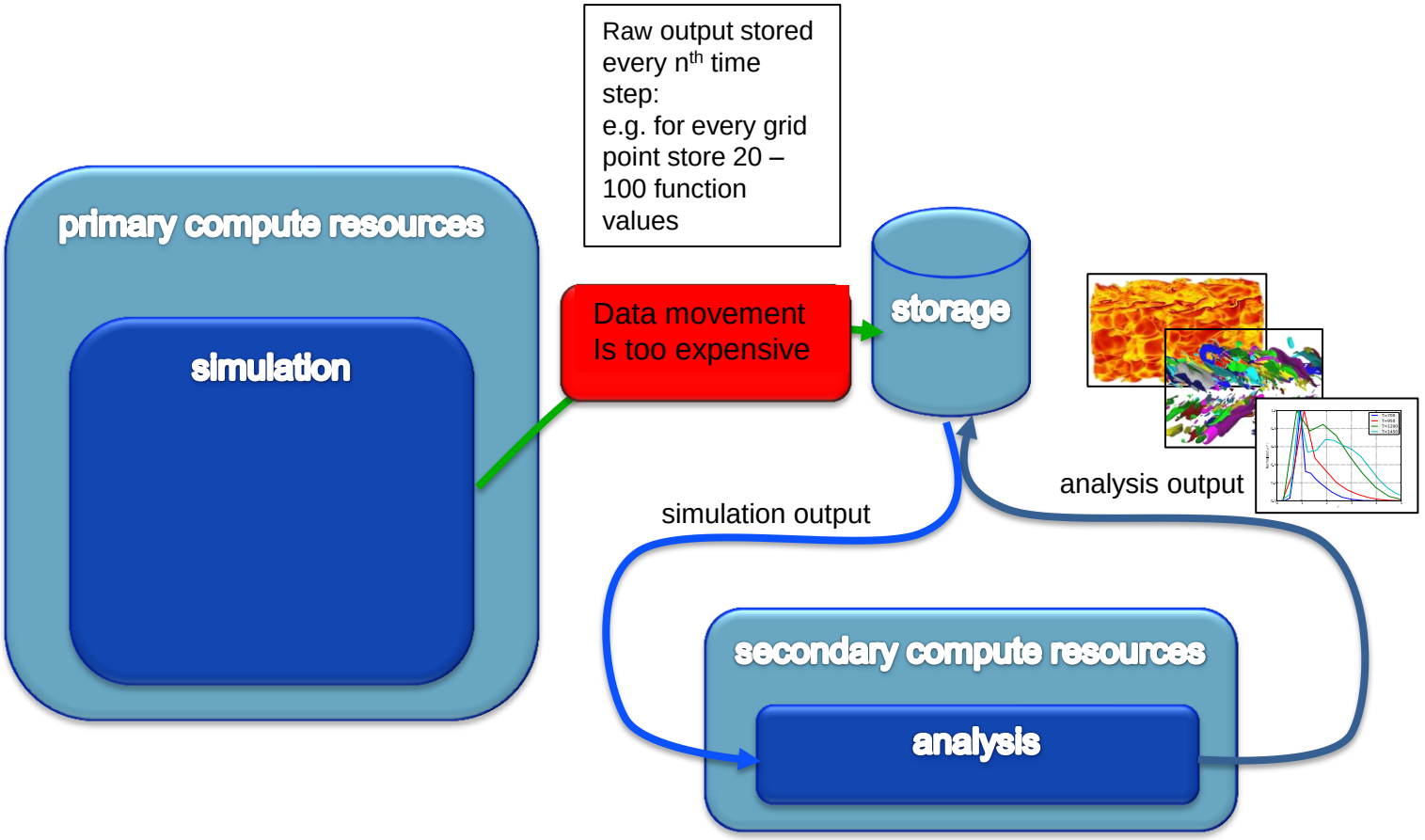
First stage

Extraction of scientific insight is a post-process on secondary resources

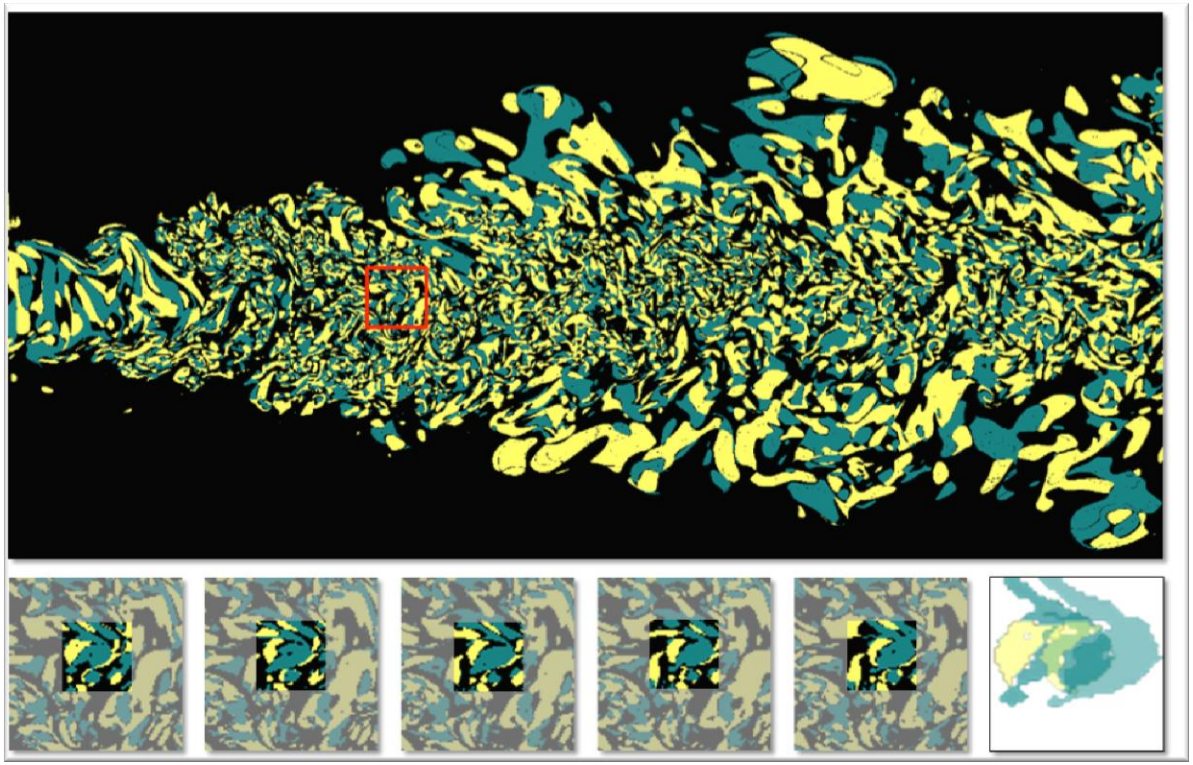


Second stage

This approach does not scale!

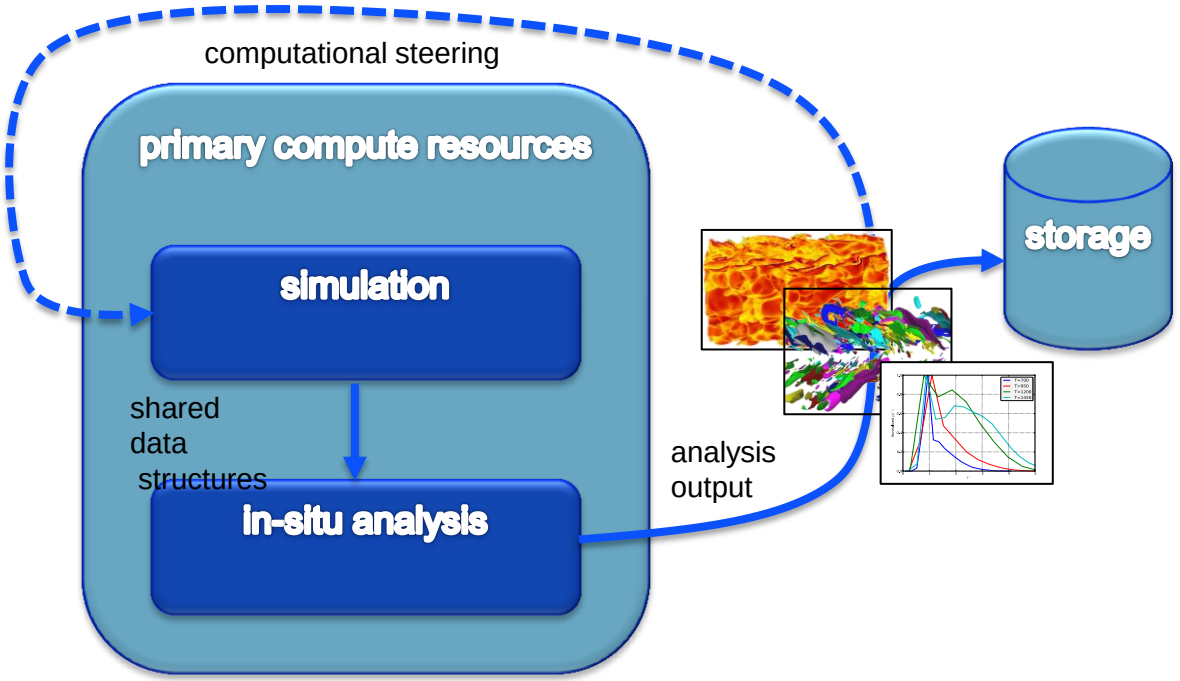


‘Science’ is lost when data is not saved to disk at sufficiently high a frequency

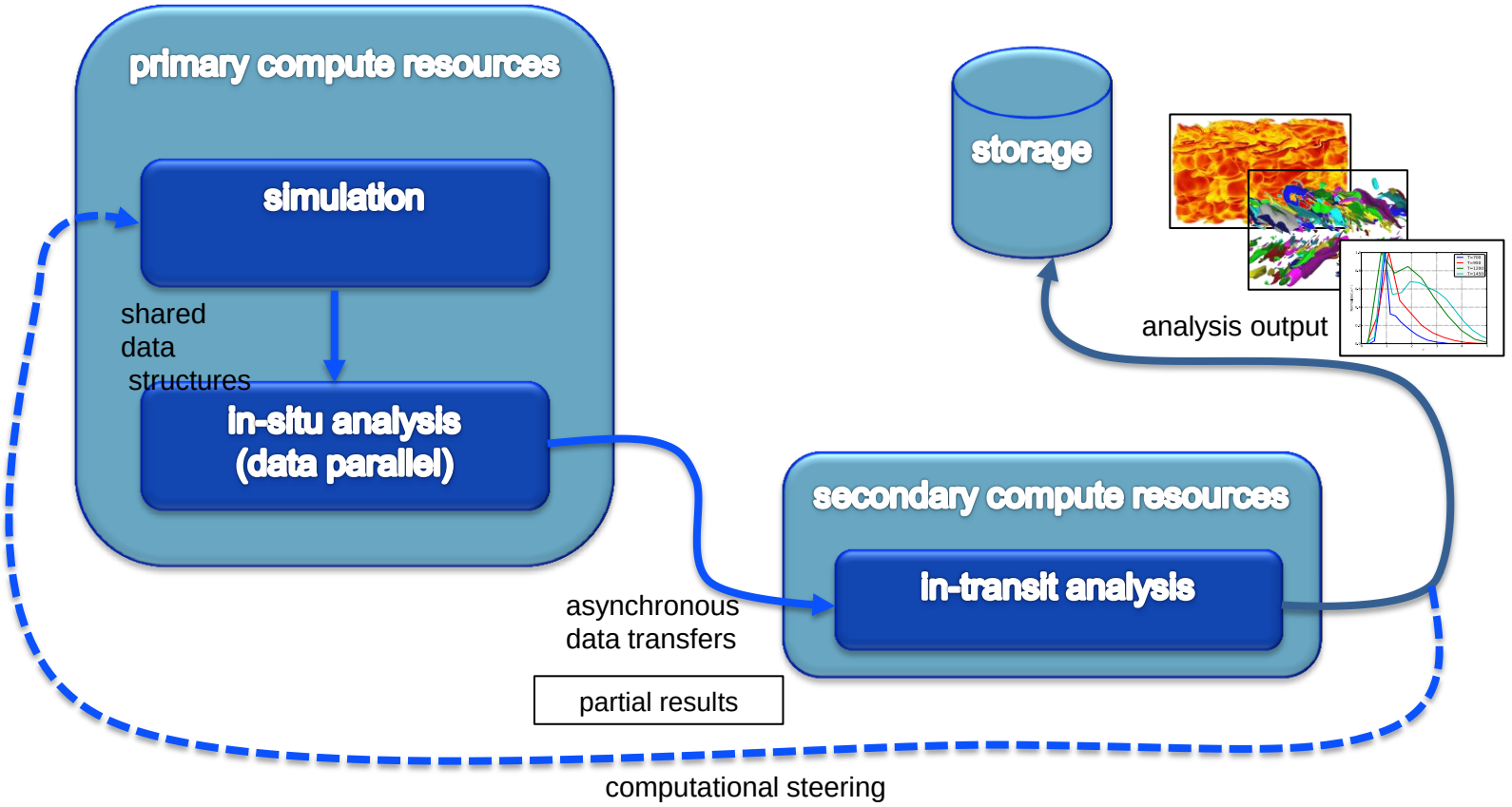


- Increased compute power allows for increased time-scale resolutions
- The additional data cannot be saved to disk due to I/O costs

To address these issues the community is shifting toward concurrent analysis



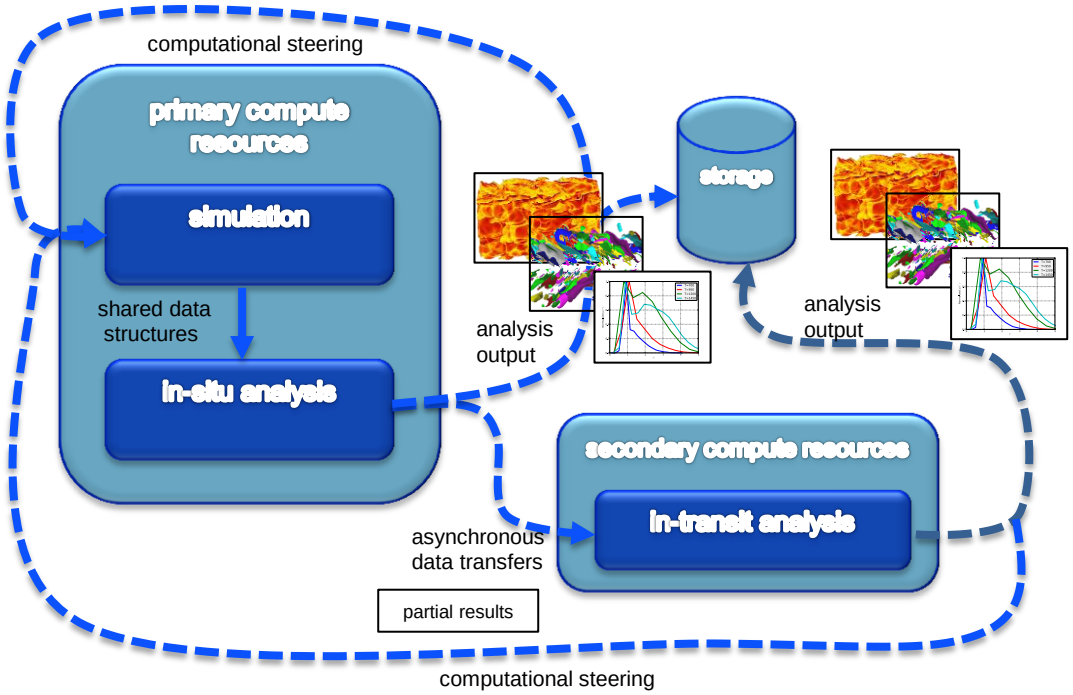
Secondary compute resources can be used to perform in-transit analysis



Sandia National Laboratories

The shift to concurrent analysis poses many R&D challenges

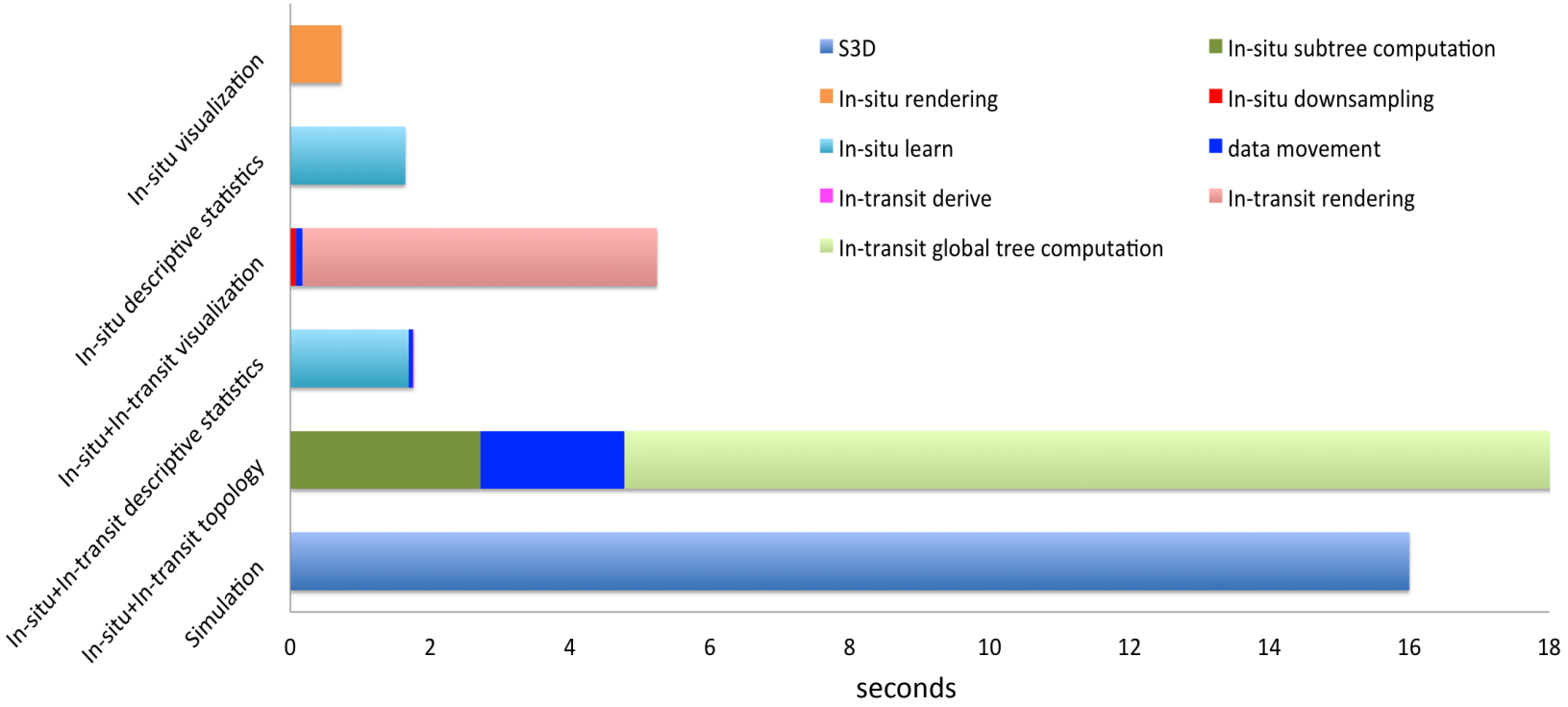
- resilient analyses
- in-situ/in-transit decomposition
- minimize communication
- shared data structures
- strict time constraints
- scheduling
- data reduction
- input parameters
- efficient data movement



Sandia National Laboratories

Hybrid in-situ + in-transit framework shows promise

timing breakdown among the simulation and analytics using 4896 cores



Combining In-situ and In-transit Processing to Enable Extreme-Scale Scientific Analysis (J. Bennett, H. Abbasi, P-T Bremer, R. Grout, A. Gyulassy, T. Jin, S. Klasky, H. Kolla, M. Parashar, V. Pascucci, P. Pebay, D. Thompson, H. Yu, F. Zhang, and J. Chen, submitted for review.)

PREPARING FOR EXASCALE

Applications work can start today.

Mantevo* Project

* Greek: augur, guess, predict, presage

- Multi-faceted application performance project.
- **Started 4 years ago.**
- Two types of packages:
  - **Miniapps**: Small, self-contained programs.
    - **MiniFE/HPCCG**: unstructured implicit FEM/FVM.
    - **phdMesh**: explicit FEM, contact detection.
    - **MiniMD**: MD Force computations.
    - **MiniXyce**: Circuit RC ladder.
    - **CTH-Comm: Data exchange pattern of CTH.**
  - **Minidrivers**: Wrappers around Trilinos packages.
    - **Beam**: Intrepid+FEI+Trilinos solvers.
    - **Epetra Benchmark Tests**: Core Epetra kernels.
    - **Dana Knoll working on new one.**
- Open Source (LGPL)
- Staffing: Application & Library developers.

Exascale presents many R&D and ‘mission’ opportunities

- **We have ideas on how to address some of the key challenges.**
- **We have an approach (co-design) that brings a holistic, integrated system engineering methodology to bare.**
- **Commercial sector is moving in the right direction by default, at least in some key areas.**
- **We can strategically accelerate critical technologies with national-scale program funding.**

We will build exascale computers, and we will figure out how to run our codes on them.


OPPORTUNITY

Acknowledgements

Sandia Collaborators: Eric Anger, Rob Armstrong, Janine Bennett, Sudip Dosanjh, Gilbert Hendry, Mike Heroux, Jackson Mayo, David Thompson

Prof Jesus Carretero and the ISPA-2012 organizers for the opportunity to address this audience

Thank You

Robert L. Clay
rlclay@sandia.gov
+1-209-610-2929