

PETASCALE FOR OPEN SCIENCE AT THE ARGONNE LEADERSHIP COMPUTING FACILITY

Ray Bair
Chief Computational Scientist
Computing, Environment and Life
Science Directorate

Argonne National Laboratory
and The University of Chicago

Fall Creek Falls Conference
September 9, 2008



ARGONNE LEADERSHIP COMPUTING FACILITY



- Fastest open science machine; 3rd fastest overall

- ▶ Peak: 557 Teraflops
- ▶ Linpack: 450.3

- Configuration

- ▶ 163,840 cores
- ▶ 80 Terabytes of memory
- ▶ 8 Petabytes of disk storage
- ▶ 10,000 volume tape archive

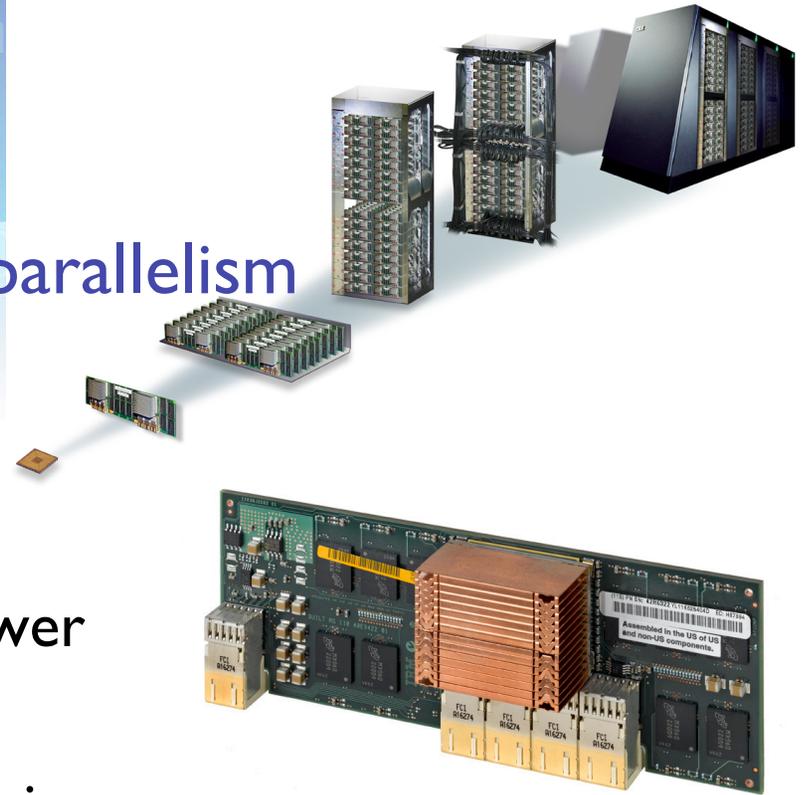


Intrepid: IBM Blue Gene/P



BLUE GENE DNA

- Low power design → massive parallelism
 - ▶ Leader in Green Computing
- System on a Chip (SoC)
 - ▶ Improves Price / Performance
 - ▶ Reduces system complexity & power
- Custom designed ASIC
 - ▶ Reducing overall part count, reducing errors
 - ▶ Permits tweaking CPU design to reduce soft errors
- Dense packaging **2.8 watts per sustained gigaflops**
- Fast communication networks
- Sophisticated RAS (reliability, availability, serviceability)
- Dynamic software provisioning and configuration



IT'S YOUR MACHINE



- Private networks
- Private I/O pathways
- Physically contiguous nodes
- Clean copy of operating system
- Repeatable performance
- Powerful options
 - ▶ Alternate compute node kernels
 - ▶ Open Source systems software

UNIQUE AND CHALLENGING FEATURES

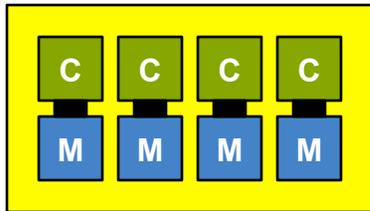


- Relatively low memory per CPU core (but very large aggregate).
 - ▶ BG/P: 2 Gbytes / 4 cores
 - ▶ True SMP is possible (sharing data structures)
- Single node optimization is important
 - ▶ Use of “double hummer” FPU requires hand tuning and experimentation
 - ▶ Strategies: Good math libraries, performance counters, code tools
- Scalable I/O strategy is required
 - ▶ One file per process **strongly** discouraged
 - Who really wants 100K files per snapshot, with **millions** per large scale run?
 - ▶ PnetCDF and HDF5 are good strategies for effectively using parallel storage system (GPFS, PVFS)
- Debugging at scale remains challenging
 - ▶ Tool groups are helping, but the issue is nevertheless hard

MPI ON BG/P – RUN MODES

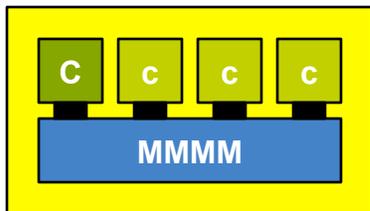


Virtual Node Mode (VNM)



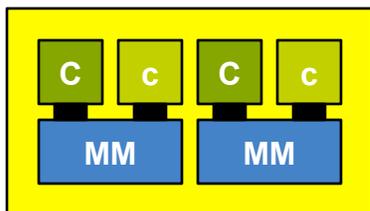
- ▶ Each core gets own MPI rank, kernel, L1-L2, 1/4 memory
- ▶ 2 L3 caches shared among 4 threads; no other threads
- ▶ CPU does both compute and communication; DMA helps
- ▶ 6 torus and 2 tree network connections are shared

SMP Mode



- ▶ Full memory available to each thread
- ▶ All resources dedicated to single kernel image
- ▶ Can start up to 4 pthreads/OpenMP threads

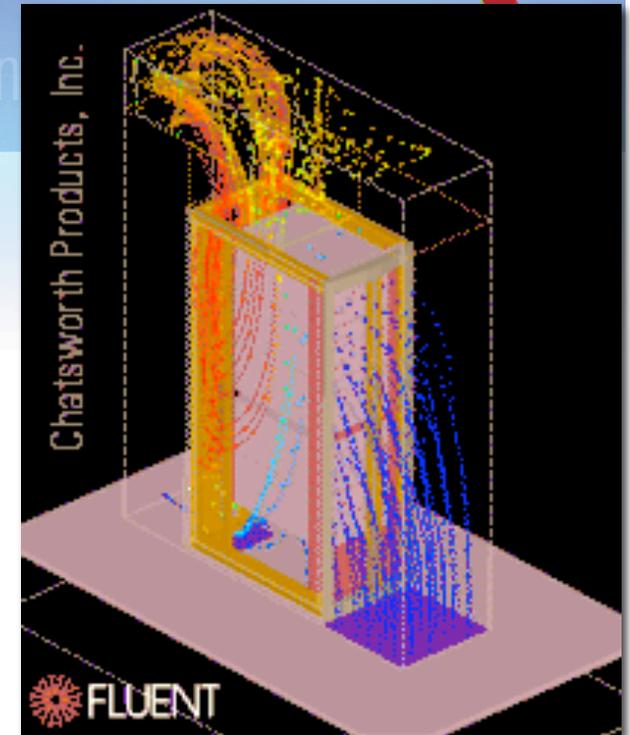
Dual Mode



- ▶ Hybrid of the two modes
- ▶ 2x MPI ranks of SMP, each can start 1 additional thread
- ▶ 1/2 memory of SMP

DATA ANALYTICS FROM GRAPHSTREAM

- Eureka (to be INCITE production)
 - ▶ (100) 2.0 GHz, 8 core, 32 GB RAM servers
 - ▶ (200) NVidia Quadro FX5600 GPUs (via S4)
 - Largest S4 installation at this time
 - ▶ Over 100 single precision TF
 - ▶ 20 KW/rack, using passive thermal management
 - ▶ Primary scratch space is the parallel file system
- Gadzooks (test and development)
 - ▶ (4) server, (8) GPU version of Eureka
 - ▶ Much broader access for test and development
- Software
 - ▶ VisIT, ParaView, vI3, VMD, etc.
 - ▶ Driven by user requirements



NVIDIA®

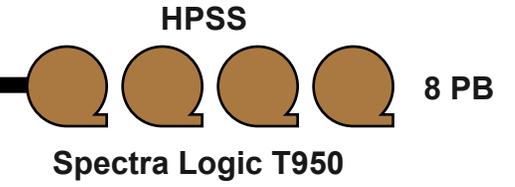
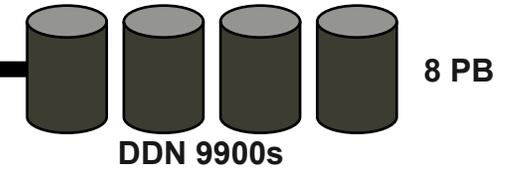
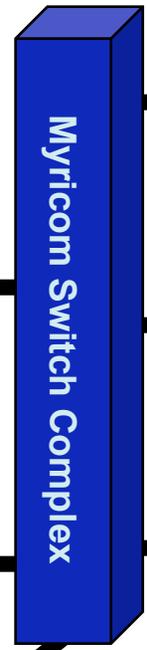
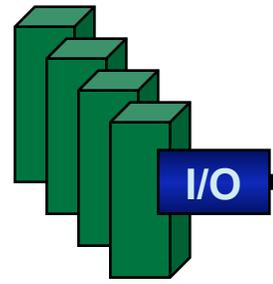
ALCF RESOURCES

INCITE

Intrepid
40 racks
160K cores
80 TB RAM
556 TF



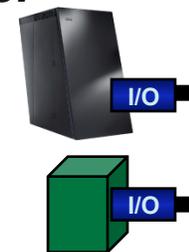
Eureka (Viz)
800 cores
50 NVIDIA S4 GPUs
100 TF



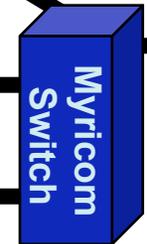
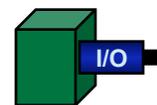
USER TEAMS
(via ESnet, UltraScienceNet,
Internet2)



Surveyor
1 rack
4K cores
13.9TF



Gadzooks (Viz)
32 cores
2 NVIDIA S4 GPUs



Test & Development

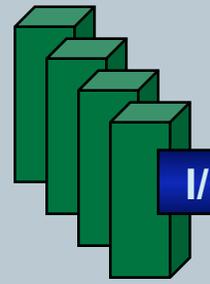
ALCF RESOURCES

INCITE

Intrepid
 40 racks
 160K cores
 80 TB RAM
 556 TF



Eureka (Viz)
 800 cores
 50 NVIDIA S4 GPUs
 100 TF

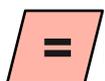
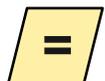
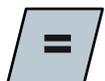


Myricom Switch Complex

DDN 9900s
 8 PB

DDN 9550s
 1.2 PB
 HPSS
 Spectra Logic T950
 8 PB

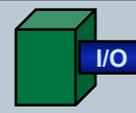
USER TEAMS
 (via ESnet, UltraScienceNet,
 Internet2)

-  = INCITE Production
-  = Early Science
-  = Being Installed

Surveyor
 1 rack
 4K cores
 13.9TF



Gadzooks (Viz)
 32 cores
 2 NVIDIA S4 GPUs

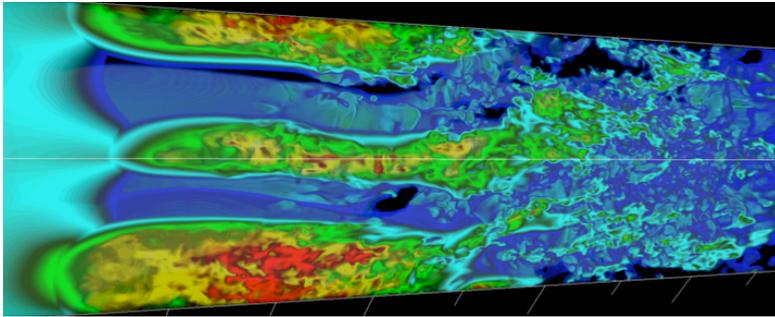


Myricom Switch

128 TB
 DDN 9550

Test & Development

Nuclear Burning (FLASH)



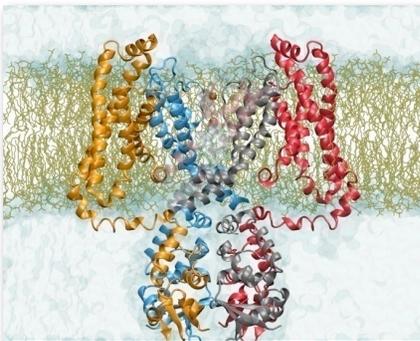
Answered critical question for buoyancy-driven turbulent nuclear burning in supernovae

Jet Engines (Pratt & Whitney)



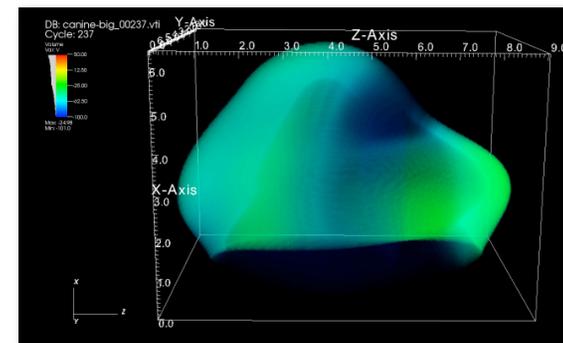
Improved designs for next generation engines that are extremely fuel efficient

Cell Biology



Showed properties of electric fields for membranes

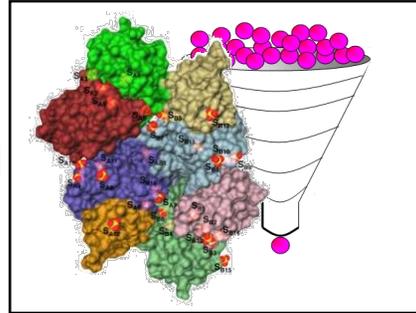
Cardiac Simulation



Study wave break and the onset of cardiac arrhythmia

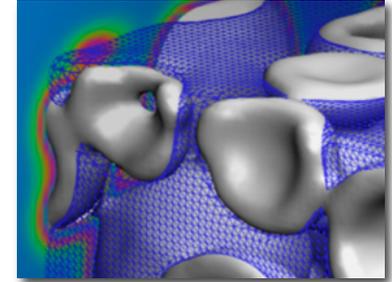
Protein Ligand Docking

Discretionary CS project to enable DOCK code for improved drug discovery



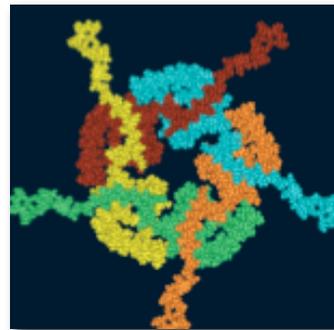
Bubble Formation (Procter & Gamble)

Improved understanding of foams and surfactants that can lead to better, safer products



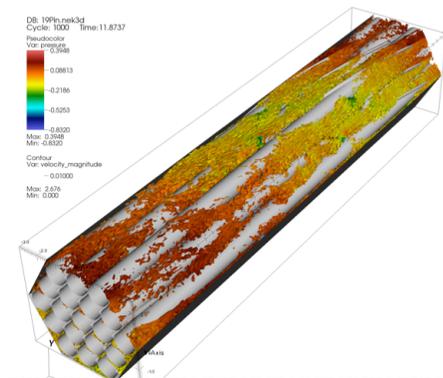
Parkinson's Disease

Provided new insights into the molecular mechanism for Parkinson's disease and its progression



Fission Reactor Design

New understanding of the effects of coolant flow on reactor designs



ALREADY RUNNING ON > 130,000 CORES DURING EARLY SCIENCE PERIOD

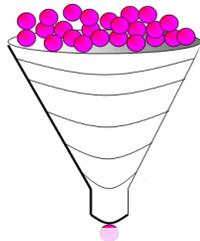
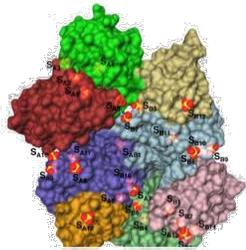


- FLASH: astrophysics /hydrodynamics
- MILC: Quantum Chromodynamics
- CPS: QCD
- Chroma: QCD
- NEK: fluid dynamics
- GTC: fusion plasma
- DOCK5+DOCK6
- QBOX: 1st principles MD
- MGDC: quantum MD
- RXFF: semi-classical MD
- GMD: classical MD
- DNS3D: turbulence
- HYPO4D: lattice Boltzmann
- PLB: lattice Boltzmann
- LAMMPS: molecular dynamics
- CACTUS: problem-solving environment

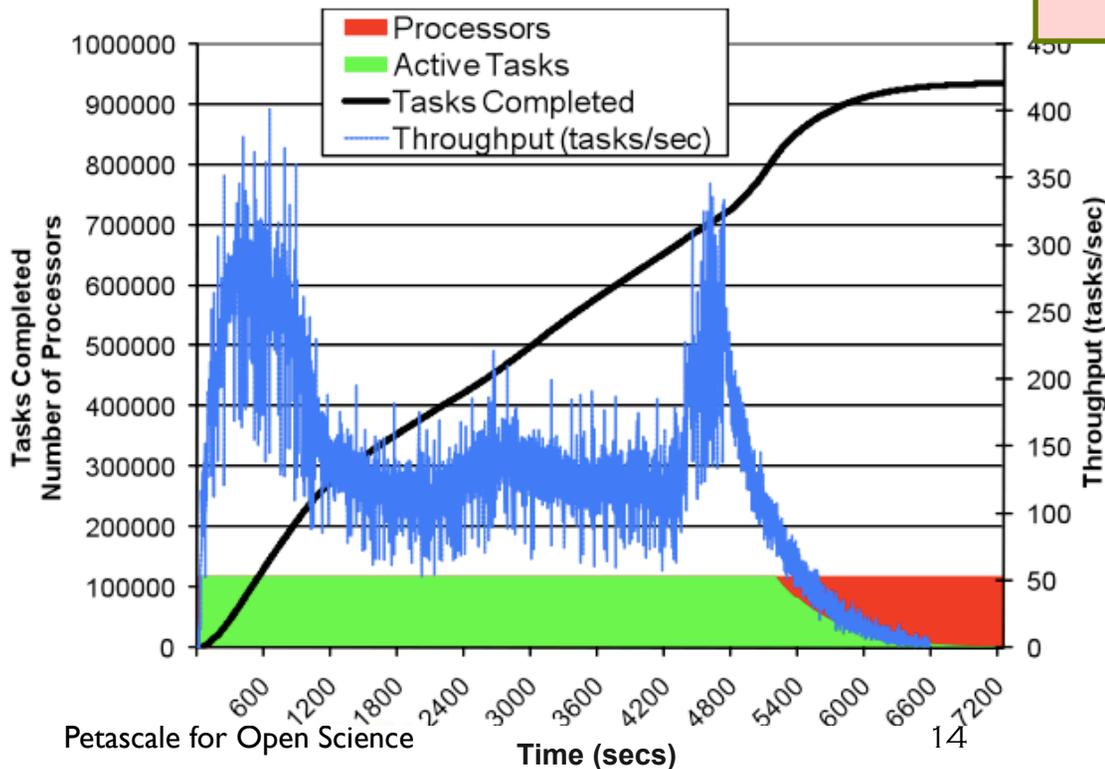
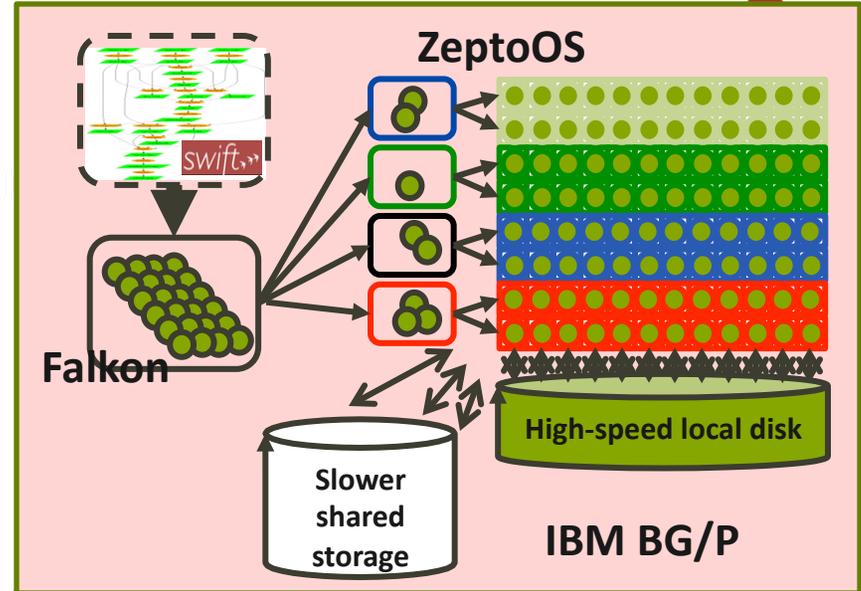
DOCK: IDENTIFYING POTENTIAL DRUG TARGETS

ZEPTOOS + FALKON + APPLICATION

Protein Target(s) x 2M+ Ligands



(Mike Kubal, Benoit Roux, and others)



CPU cores: 118,784

Tasks: 934,03

Elapsed time: 2.01 hours

Compute time: 21.43 CPU years

Average task time: 667 sec

Relative Efficiency: 99.7%

(from 16 to 32 racks)

Utilization:

■ Sustained: 99.6%

■ Overall: 78.3%

ALCF RELATED EVENTS AT SC08



- Petascale Computing Experiences on BG/P,
Wed. 11/19, 12:15 - 1:15pm.
- Blue Gene System Management Community BOF”
Tue. 11/18, 5:30 - 7:00pm.
- Blue Gene Consortium Meeting, time TBD
- SPEC MPI2007 - A benchmark to measure MPI
application performance, Tue. 11/18, 5:30 - 7:00pm

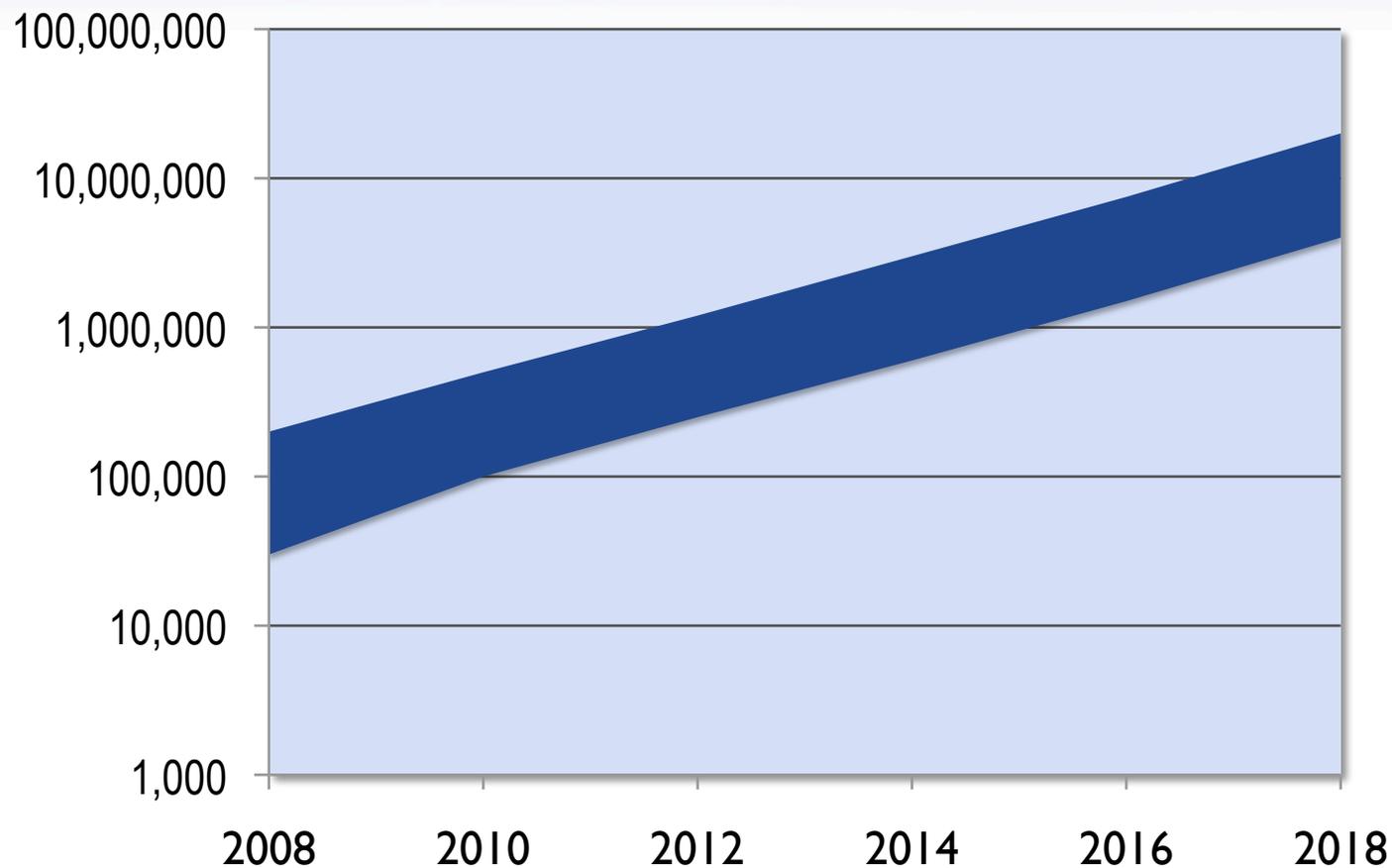


PANEL COMMENTS

**CODE DEVELOPMENT IS HARD WORK.
PLAN FOR FUTURE SYSTEMS.**



Concurrency in Leadership Systems



SOURCES OF HARDWARE CONCURRENCY



- Cores
- Threads
- FPUs, GPUs, SPUs, other coprocessors (FPGAs)
- Memory channels
- Communications and I/O channels