Cray XT4: An Early Evaluation for Petascale Scientific Computing

Jeffery Kuehn

Sadaf Alam

Richard Barrett

Mark Fahey

Jeffrey Larkin

Ramanan Sankaran

Patrick Worley

Acknowledgements

This research used resources of the National Center for Computational Sciences at Oak Ridge National Laboratory, which is supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC05-00OR22725.

Additional Thanks to: K. Goto at TACC, and for HPL: T. Chan, E. D'Azevado, K. Wong, and R. Wong, Steve Poole (ORNL) for NAMD test cases, Trey White (ORNL), Climate and weather modeling group (ORNL), Fred Jaeger et. al. (ORNL)

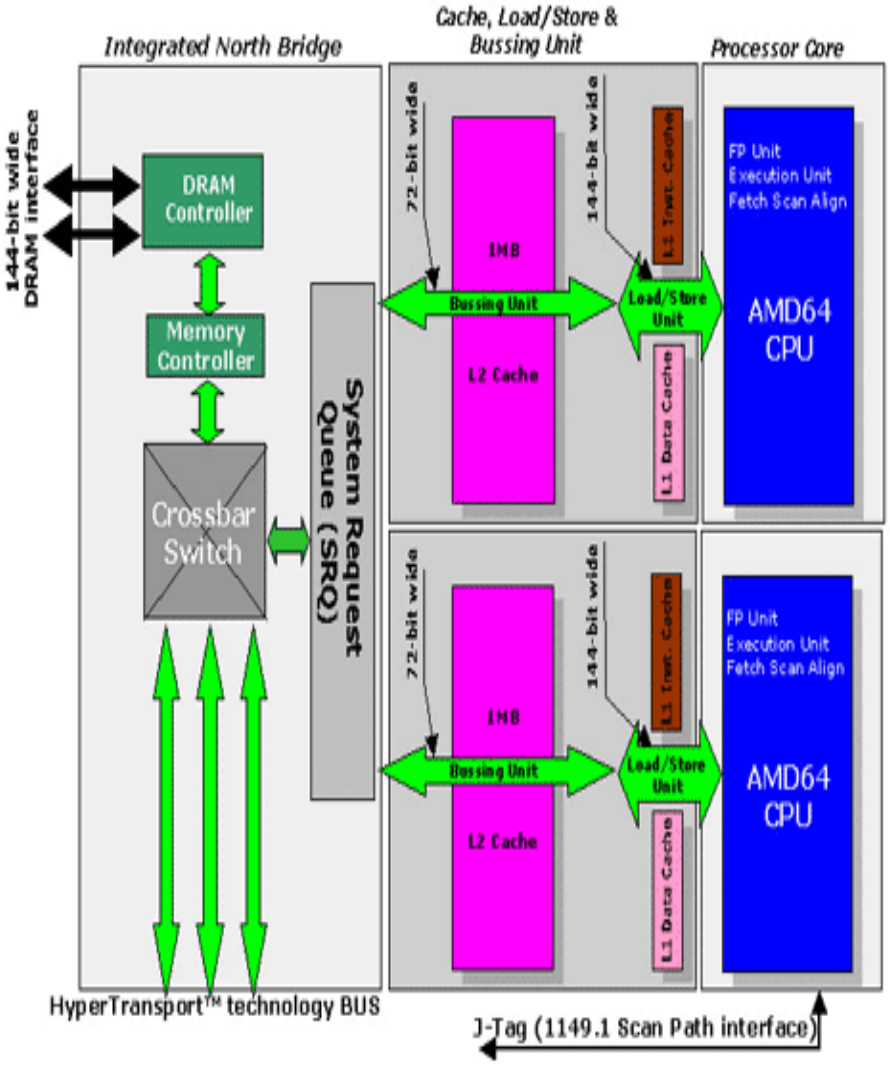
Introduction

- **NCCS Path to Petascale computing**
  - **5,212 2.4 GHz single-core nodes (XT3)**
  - **5,212 2.6 GHz dual-core nodes (XT3)—upgrade**
  - **6,296 2.6 GHz dual-core nodes (XT4)—addition**
  - **119 TFLOPS →*250 TFLOPS →1 PFLOP***

- **Evaluate impact of system evolution on critical apps**
  - **multi-core technology**
  - **custom network infrastructure**

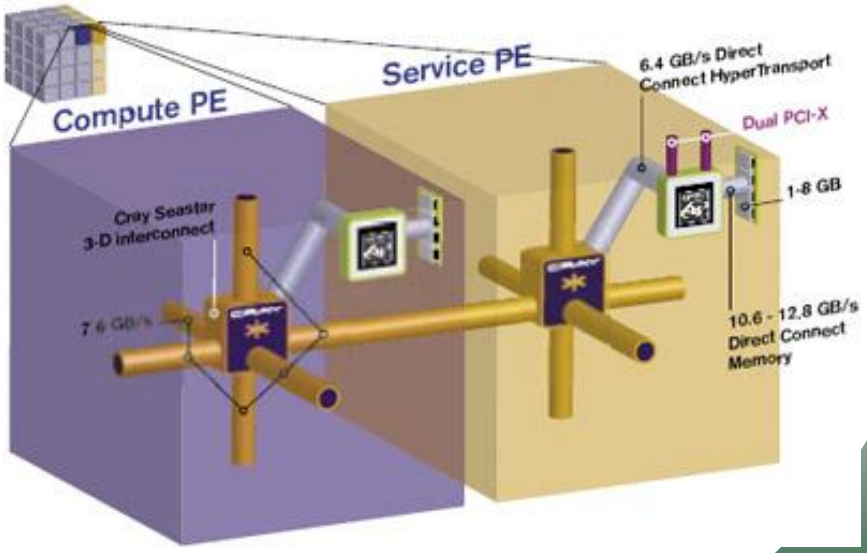
- **Cray XT systems**
  - **AMD Opteron processing nodes**
  - **3D toroidal network with SeaStar router chips**
  - **Light-weight custom OS: Catamount →*Compute Node Linux (CNL)***
  - **Lustre file system**

Overview

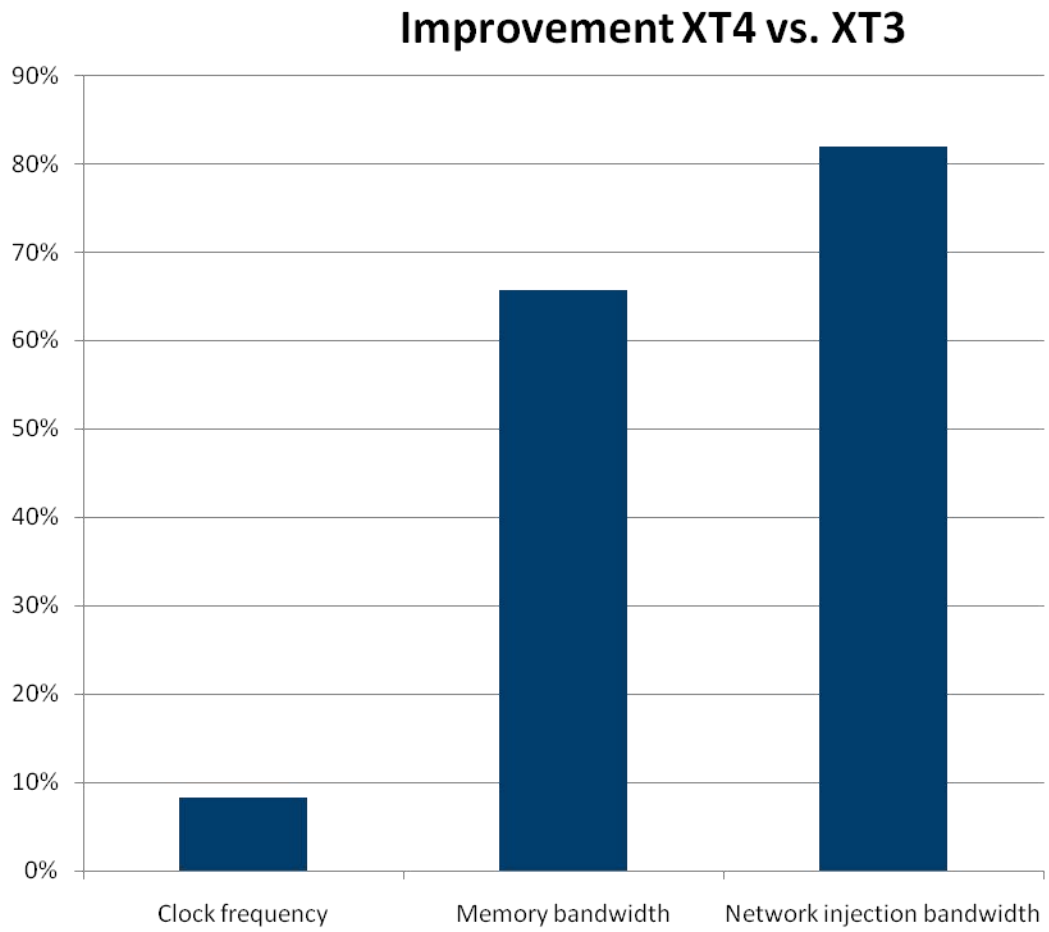


- **What has changed from XT3 to XT4**
  - **Processor: 2.4 GHz single-core -> 2.6 GHz dual-core**
  - **Memory: DDR-400 -> DDR2-667**
  - **SeaStar version 1.2 -> 2.1 improves injection bandwidth**

Cray XT4 Scalable Architecture

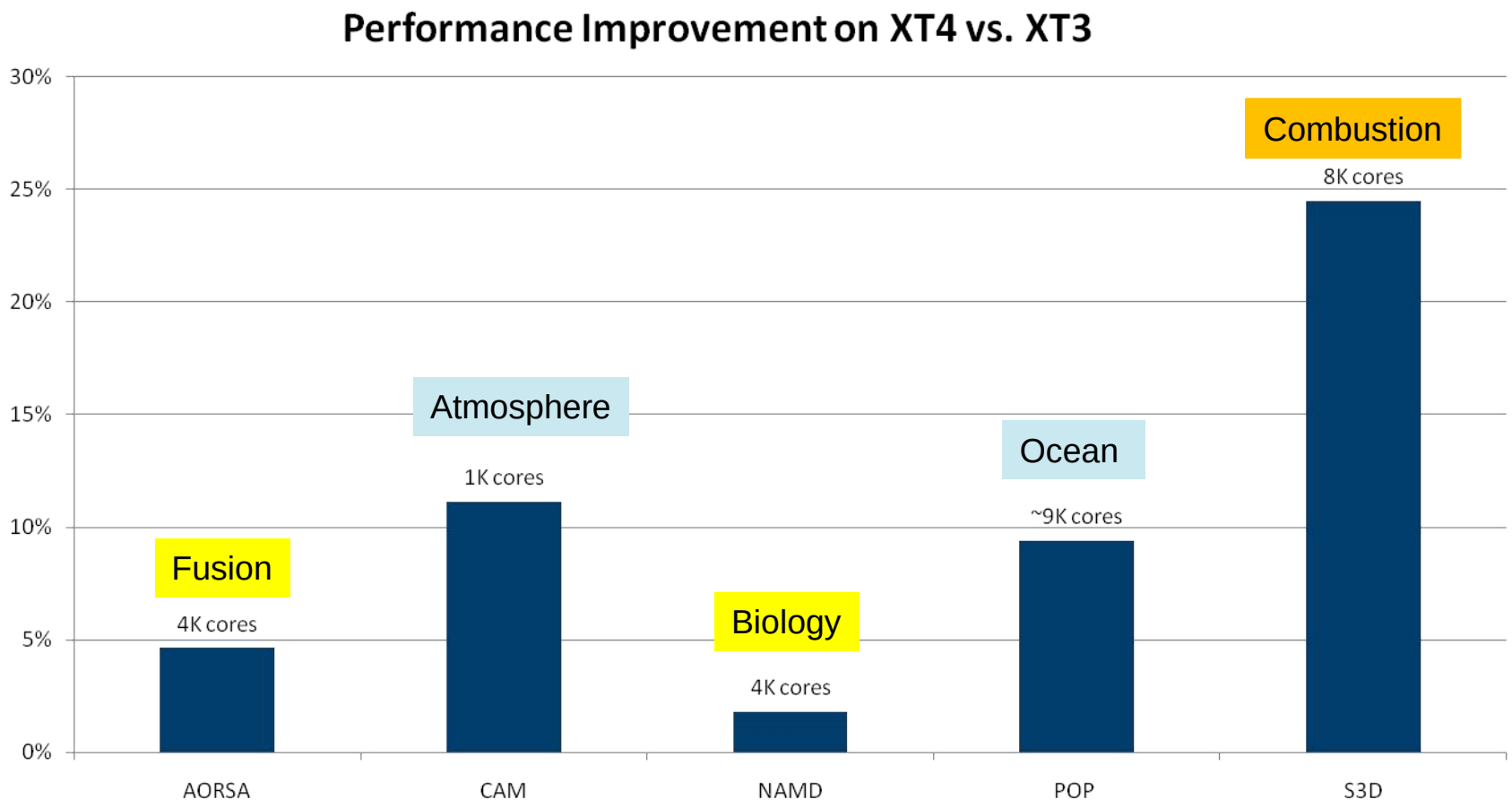


Comparison of Architectural Features



- AMD Rev E Opteron vs. Rev F Opteron (upgradable to quad-core)
- DDR vs. DDR2 support
- SeaStar vs. SeaStar2

Snapshot of Early Application Performance



List of results in the paper

- **Micro-benchmarks**
  - **HPCC**
    - **Network latency and bandwidth**
    - **FFT, DGEMM, RA and Streams**
    - **HPL, MPI-FFT, PTRANS and MPI-RA**
  - **Bi-directional MPI bandwidth**
- **Applications**
  - **CAM (3 sets of results)**
  - **POP (3 sets of results)**
  - **NAMD (2 sets of results)**
  - **S3D**
  - **AORSA**

OAK RIDGE National Laboratory

Outline

- Characterization using micro benchmarks
- Application scaling
- Conclusions and future plans

Experimental Configurations

	XT3	XT3 Dual-Core	XT4
Processor	2.4 GHz single-core Opteron	2.6 GHz dual-core Opteron	2.6 GHz dual-core Opteron
Processor Sockets	5,212	5,212	6,296
Processor Cores	5,212	10,424	12,592
Memory	DDR-400	DDR-400	DDR2-667
Memory Capacity	2GB/core	2GB/core	2GB/core
Memory Bandwidth	6.4 GB/s	6.4 GB/s	10.6 GB/s
Interconnect	Cray SeaStar	Cray SeaStar	Cray SeaStar2
Network Injection Bandwidth	2.2 GB/s	2.2GB/s	4GB/s

Nomenclature

- Socket (Node)
- Cores per socket
- MPI task
- VN (two MPI tasks per dual-core mode
- SN (one MPI task per dual-core node)

OAK RIDGE National Laboratory

Overview of HPCC Benchmarks

Execution modes

- Global
- Embarrassingly parallel (EP)
- Single processor (SP)

Global performance

- High Performance Linpack (HPL)
- Matrix Transpose (PTRANS)
- MPI Fast Fourier Transform (FFT)
- MPI Random Access (RA)

Processor/memory performance

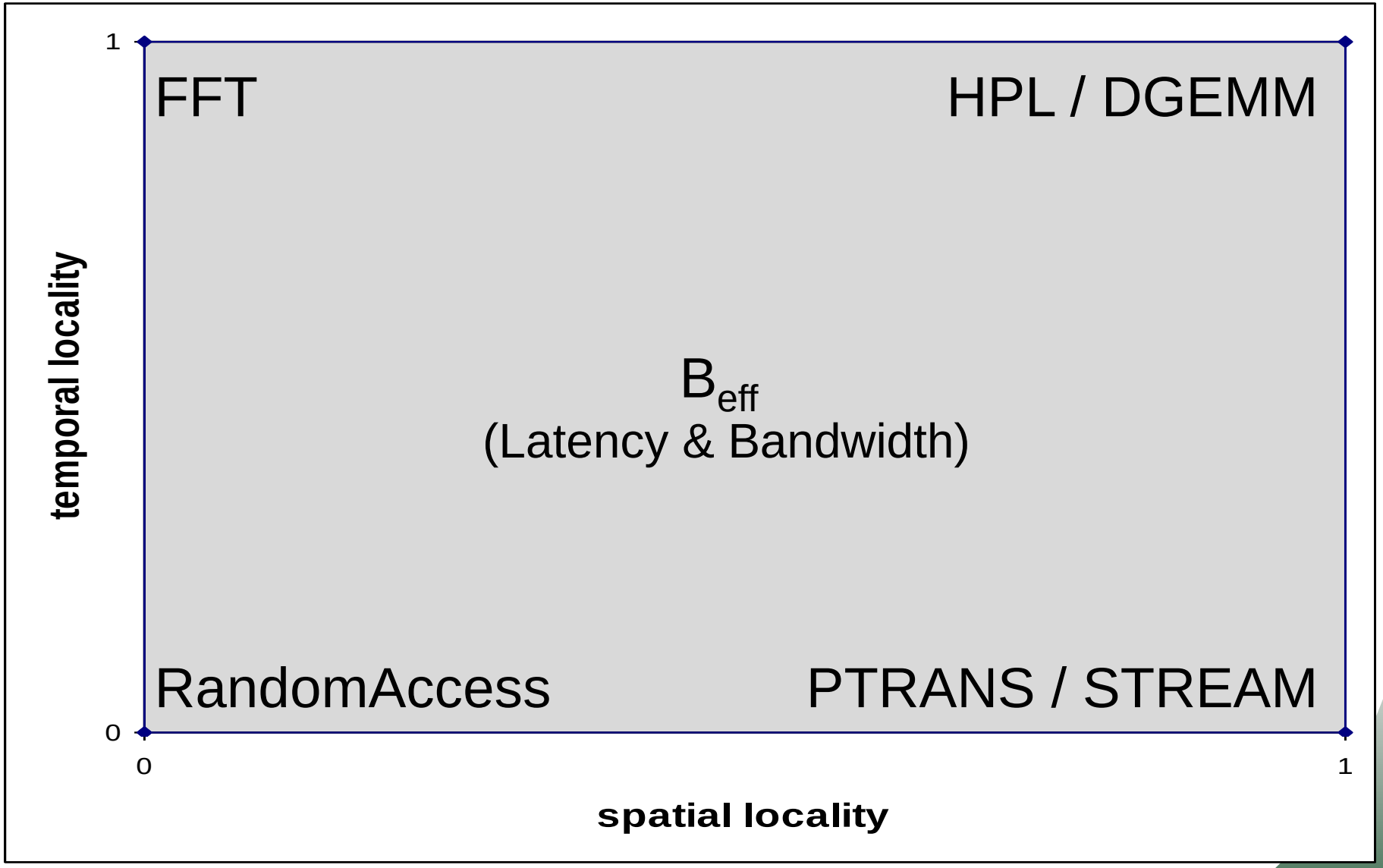
- Double-precision Matrix Multiply (DGEMM)
- STREAM
- FFT
- RA

Network performance

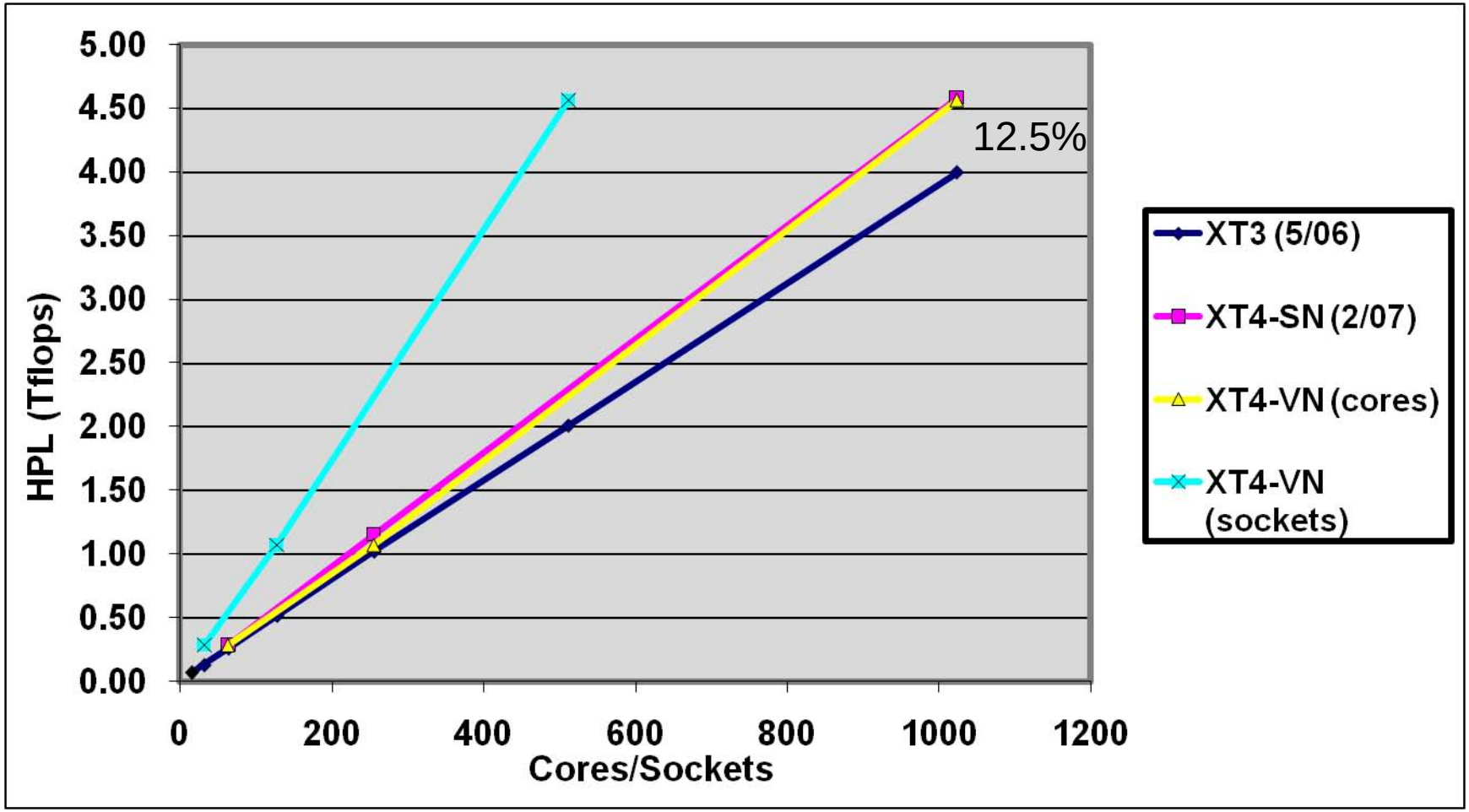
- PingPong (min-avg-max)
- Natural Ring
- Random Ring

OAK RIDGE National Laboratory

Locality View

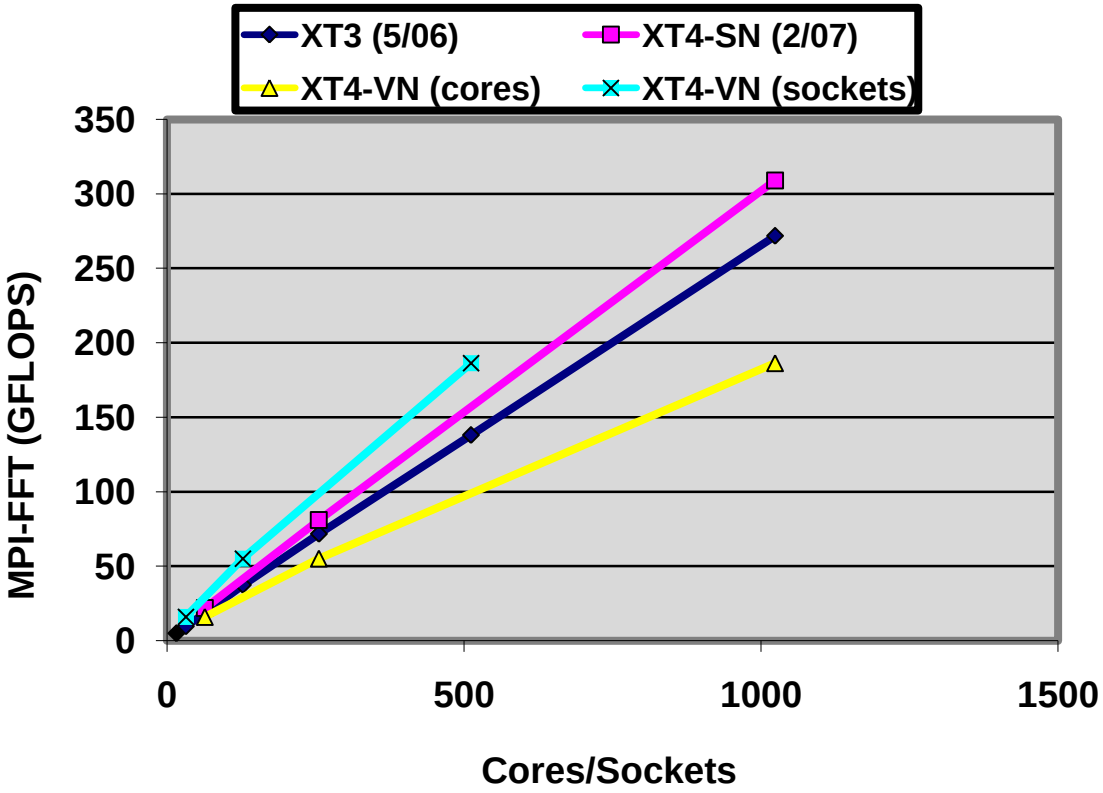


Global Performance (HPL)



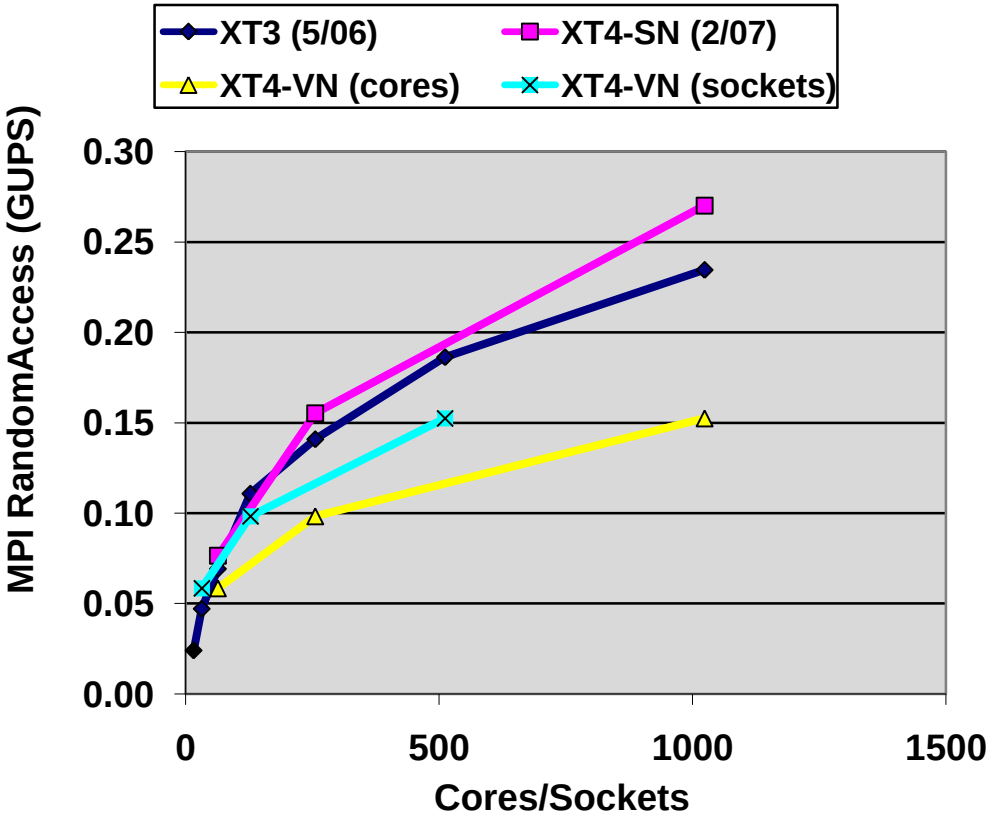
MPI FFT

- Temporal locality emphasized
  - Not as good as HPL
  - But still performs well
- Global communication
  - Network impact
    - Higher latency
    - Lower per-core bandwidth
- Multi-core advantageous



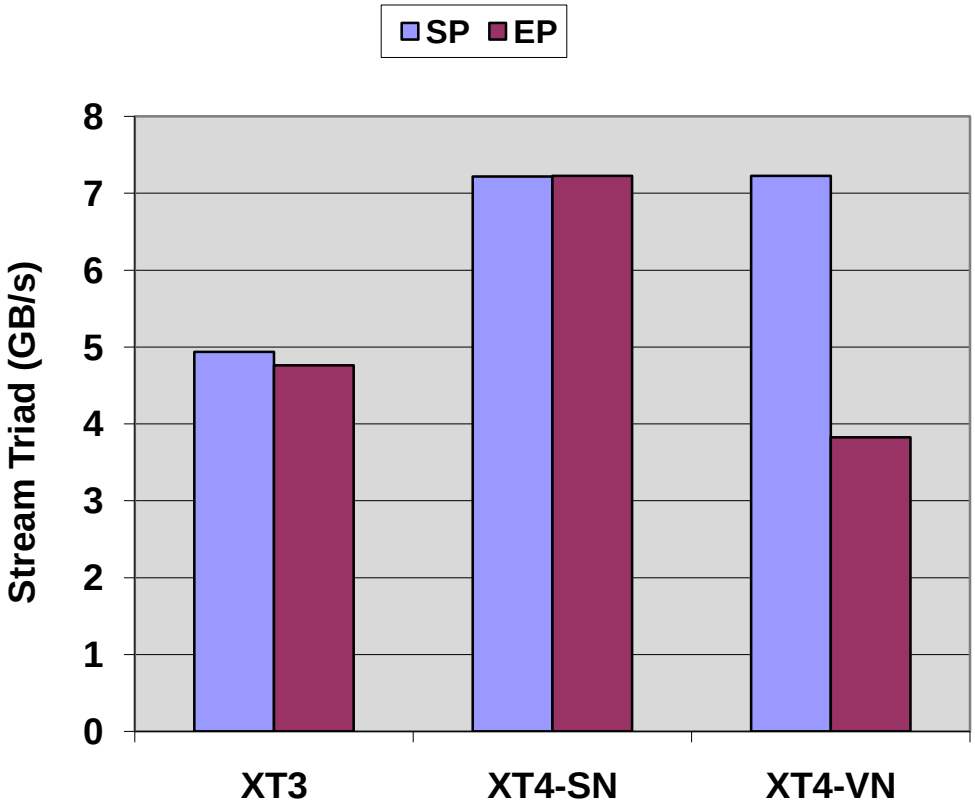
Global (MPI Random Access)

- Low locality
  - If the cache hit rate isn't zero, the test case isn't big enough ☺
- Note impact of DDR2 memory
- Comparing SP vs EP
  - Memory controller is the bottleneck
  - Engaging second core provides no benefit
  - Good multi-core code should not look like this
  - But most code does ☺



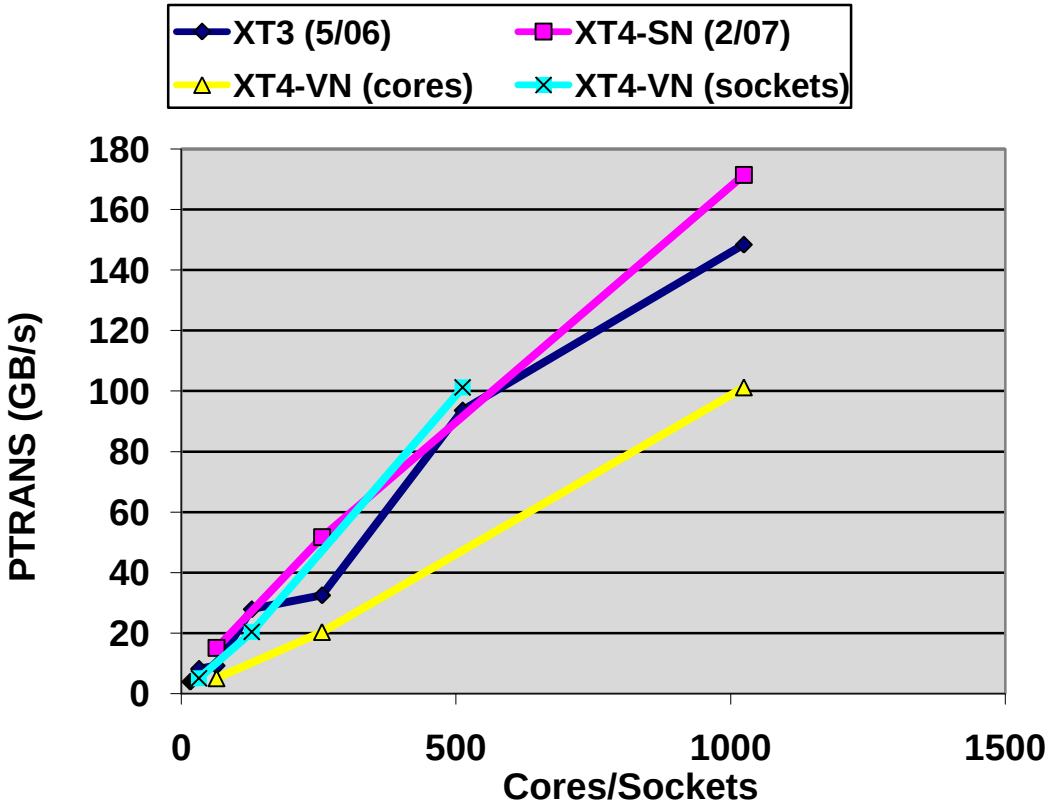
SP/EP STREAM

- Emphasizes spatial locality
- Faster DDR2 memory provides a distinct advantage over XT3
- Shared memory controller – dual (dueling?) cores see half bandwidth
- EP-Scale (best performer) shows one core can saturate memory interface
- Tip: best performance was achieved with less aggressive prefetch (9 on SN vs 8 on VN)
  - Prefetch bottleneck at memory controller?



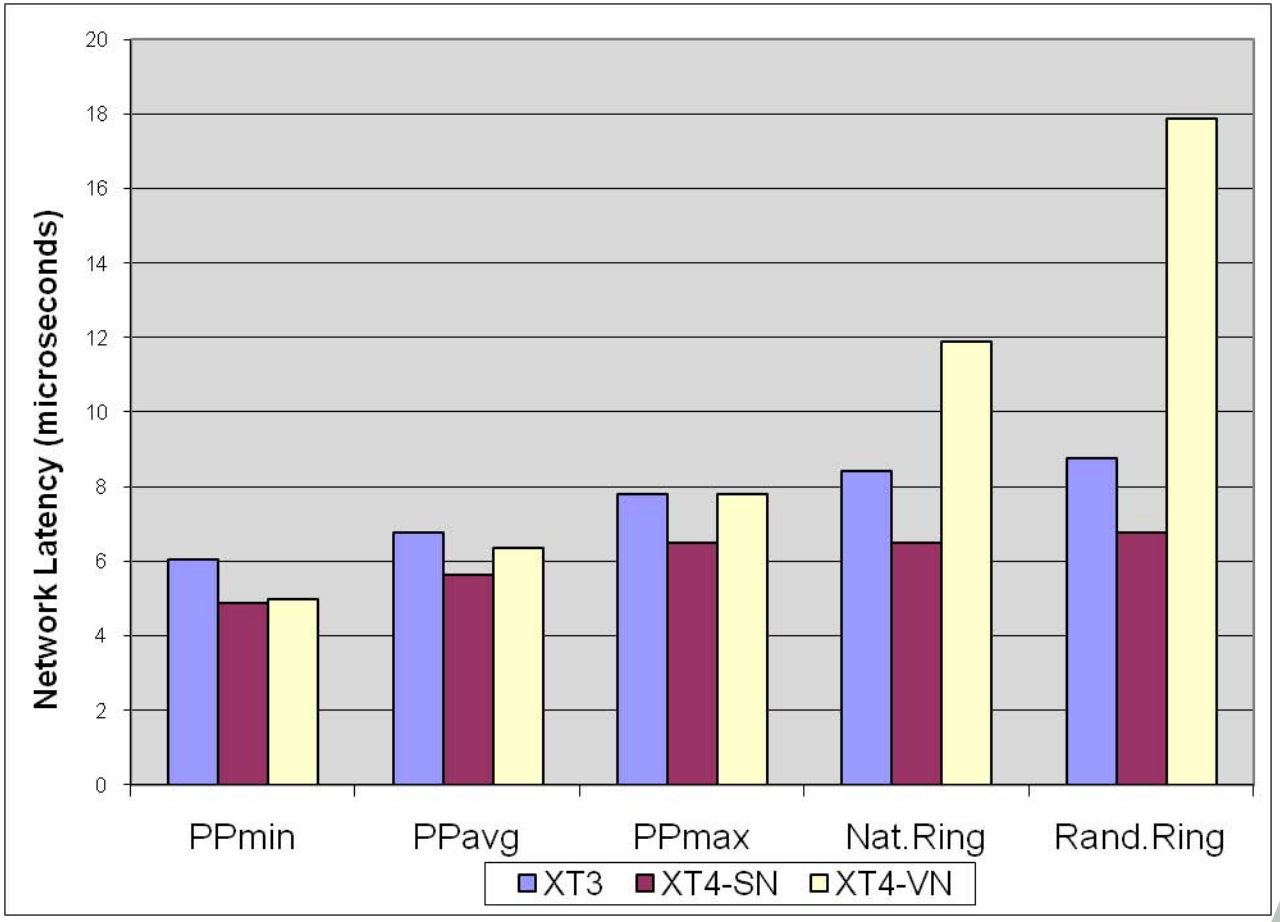
PTRANS

- Spatial locality
- Impacted by network contention
- Link BW unchanged
- Layout Impact
  - CUG 2006 PSC paper showed 10% impact on XT3



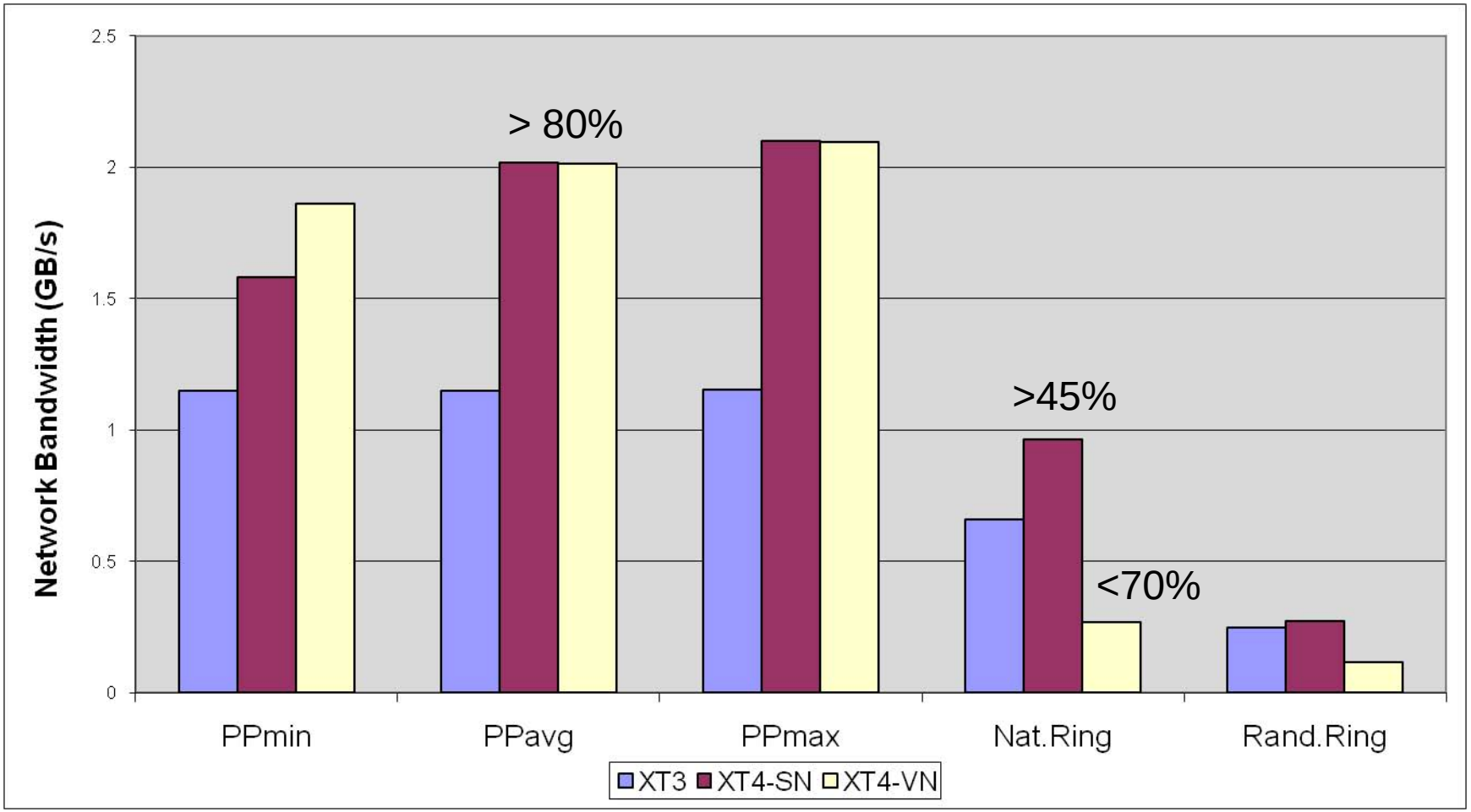
Communication (Latency)

- **Latency generally increases as core count increases – more hops**
- **VN latencies spread higher**
  - **NIC contention between cores**
  - **higher injection BW induces additional link Contention**



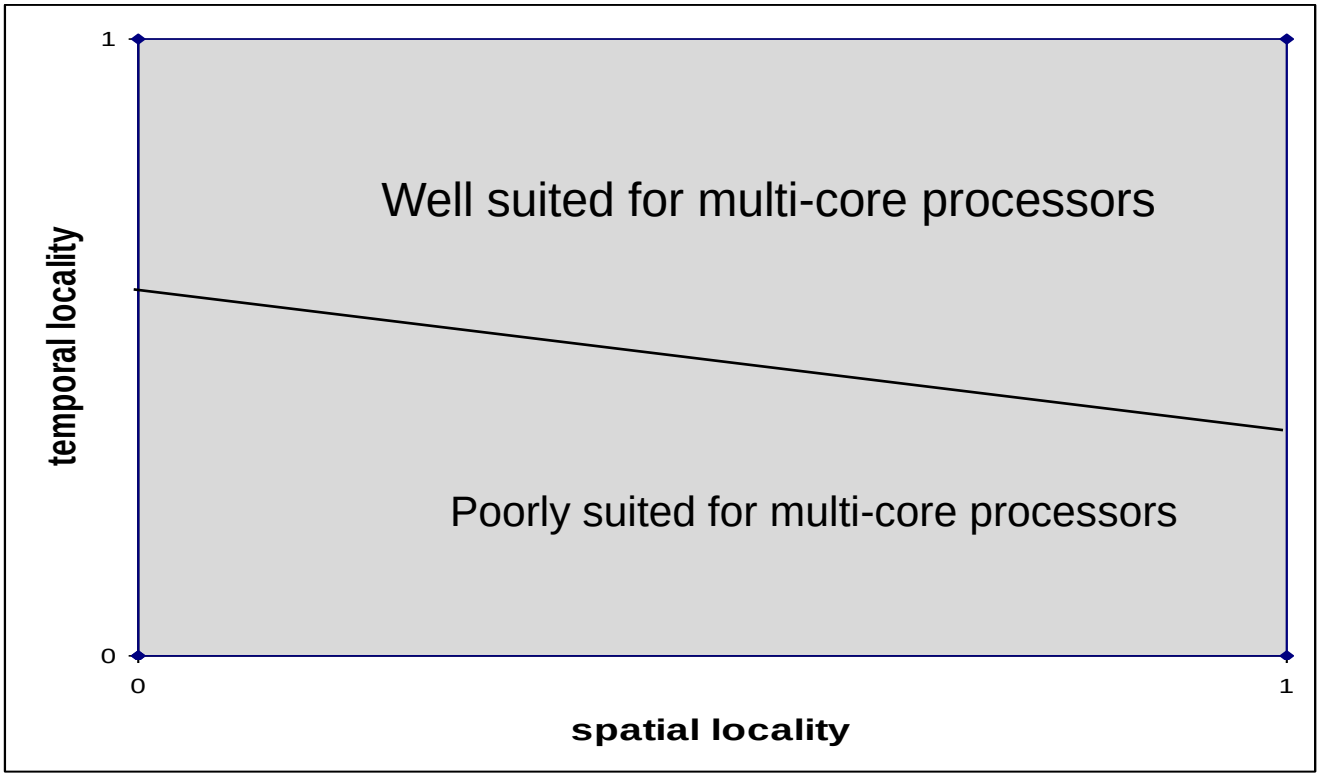
PP=PingPong (min,avg,max)
Nat.Ring=Natural Ring
Rand.Ring=Random Ring

Communication (Bandwidth)



Summary of HPCC Results

- **Impact of resource contention**
- **Outliers due to the maturity of software stack**
- **Mapping onto multi-core MPP (SP/EP results)**



Targeted Applications

SCIENCE PRODUCTIVITY ON REAL CODES

- Petascale-candidate applications and test cases
- Full physics, production science codes, in production configurations

Challenges

- Spatial and temporal locality characterization
- Complex interactions: compiler, memory, …

Goal

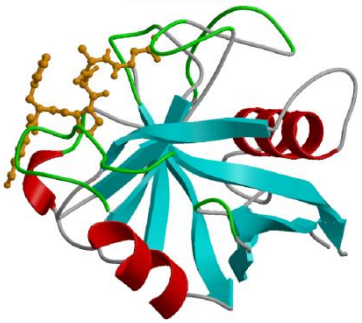
- Understand scientific productivity on this machine

OAK RIDGE National Laboratory

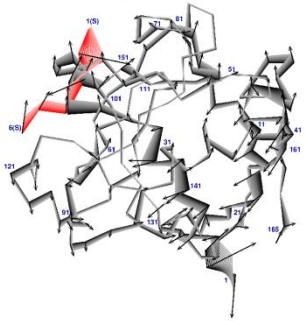
Computational Biology

- **Molecular dynamics (MD) methods:**
  - **Explicit solvent**
  - **Cut-off radius**
  - **Particle Mesh Ewald (PME)**
  - **FFT**
- **NAMD**
  - **Charm++ execution model**
  - **MPI interface**

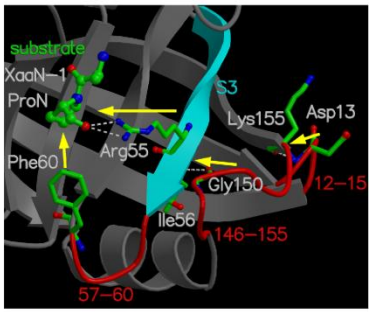
Structure



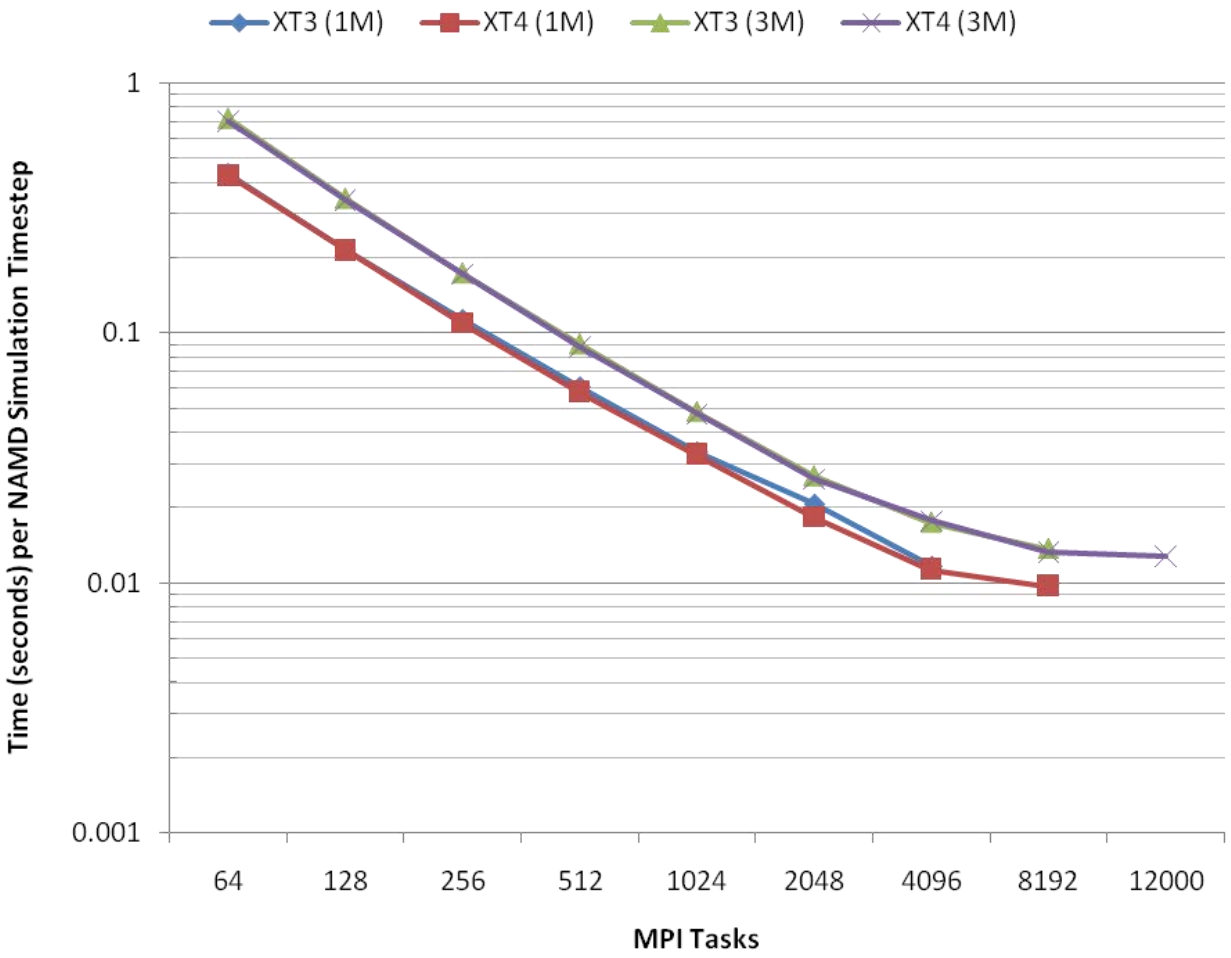
Dynamics



Function



Biomolecular Simulations (NAMD)



Climate Modeling—Atmosphere and Land Components (CAM/CLM)

- Dynamics and physics phases
- Hybrid programming model
  - MPI and OpenMP
- Scaling runs to 960 MPI tasks
- Scaling and VN efficiency of:
  - Global sum calls
  - Load balance MPI_Alltoallv calls

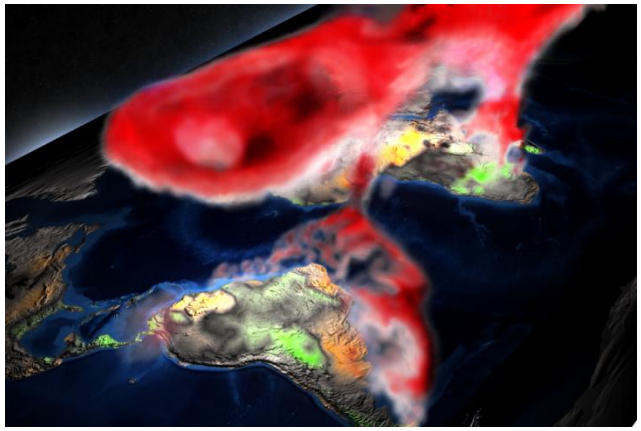
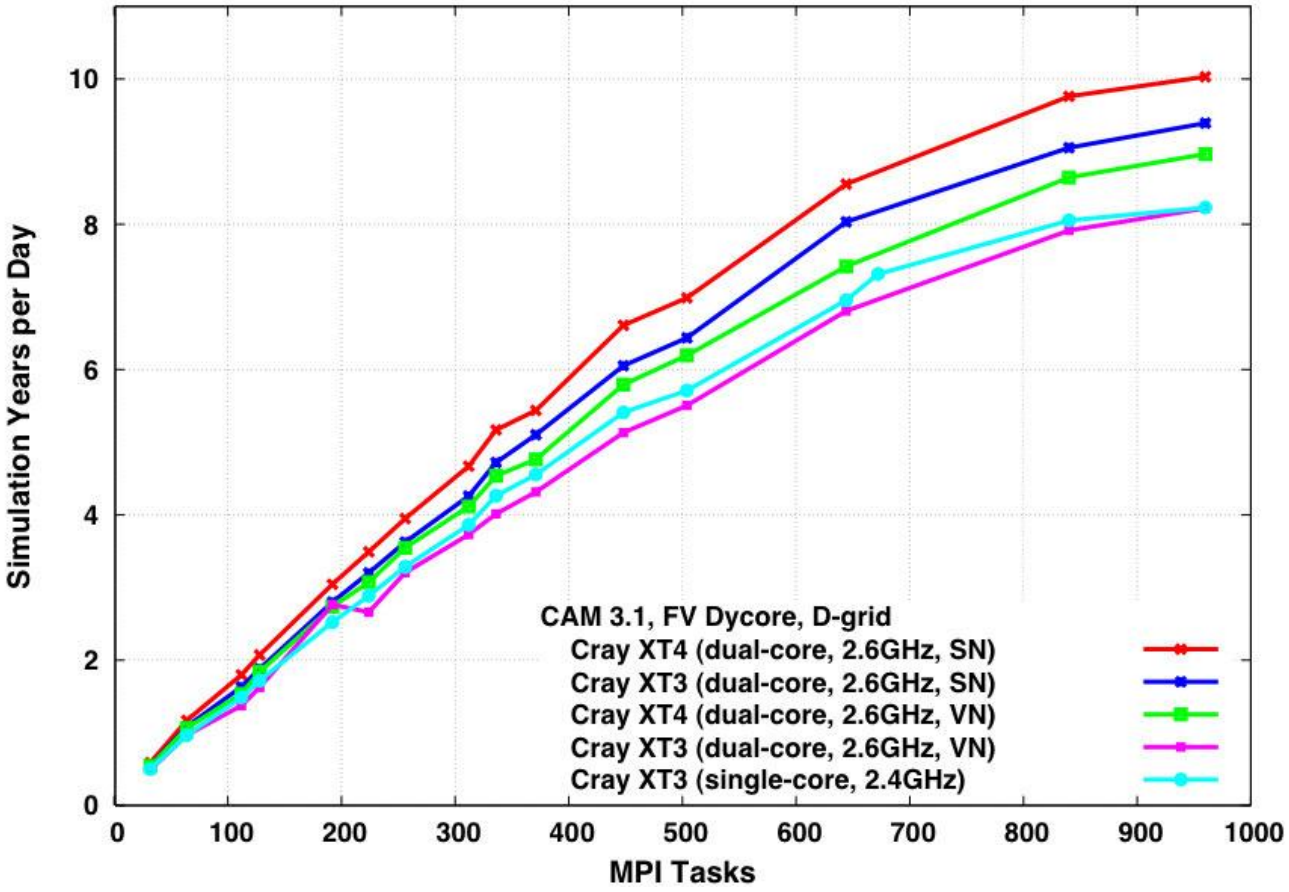


Image generated by Jamison Daniel (ORNL), Forest Hoffman (ORNL) and John Drake (ORNL)

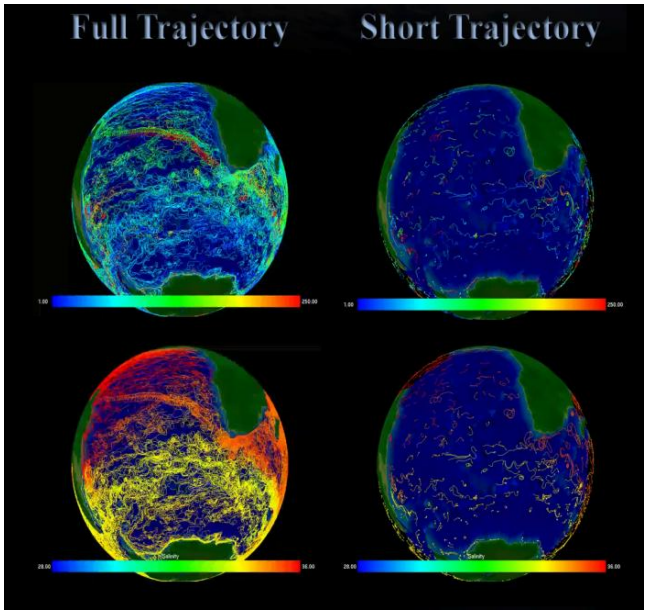
Community Atmosphere Model (CAM)

D-grid Benchmark: Finite Volume dycore
361x567x26 with 2D decomposition



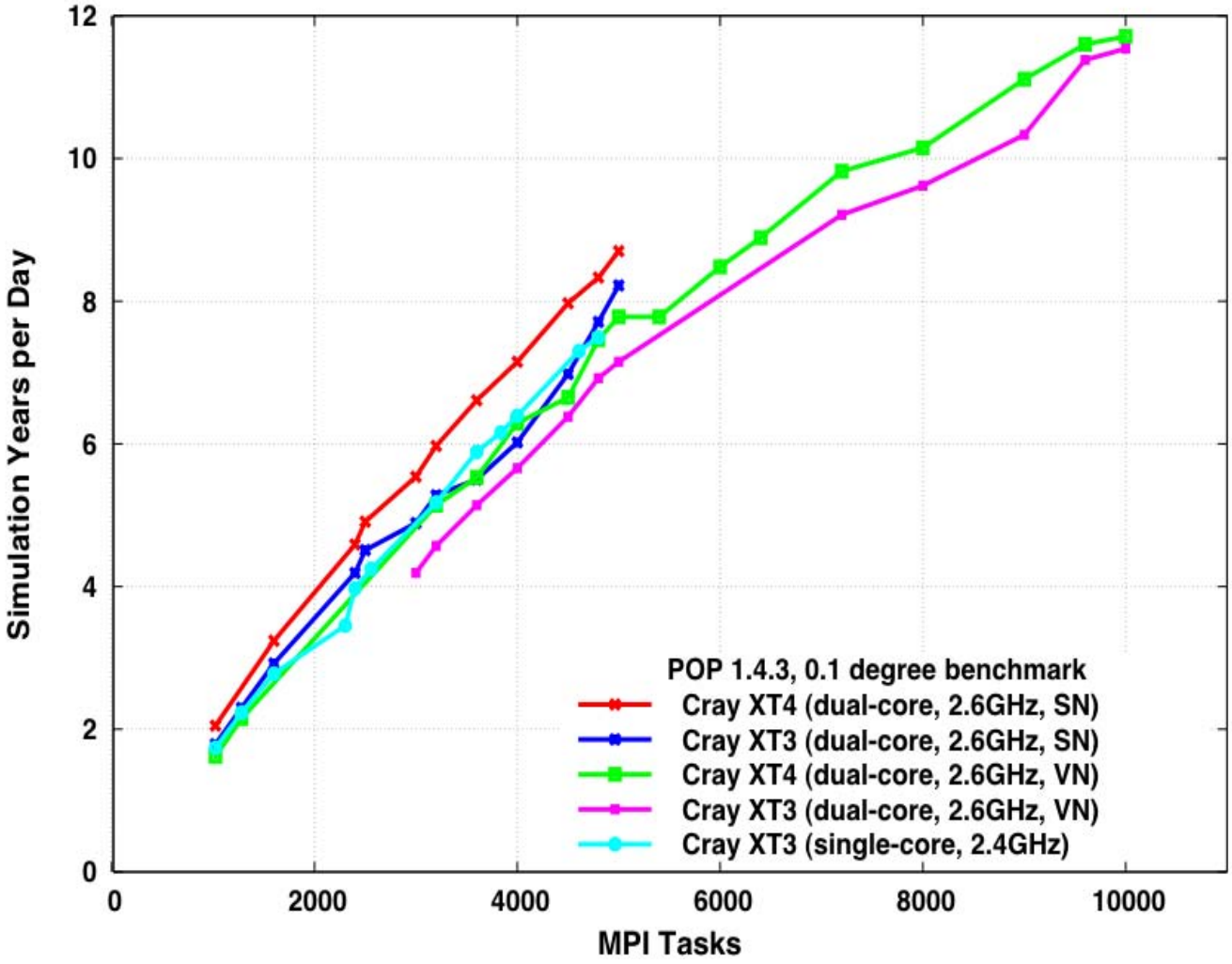
Parallel Ocean Program (POP)

- Global ocean circulation model (CCSM)
- 3D finite difference formulation
- Baroclinic and Barotropic phases
- 1/10th degree benchmark (3600x2400x40)
- Benefited from MPI_Allreduce optimization

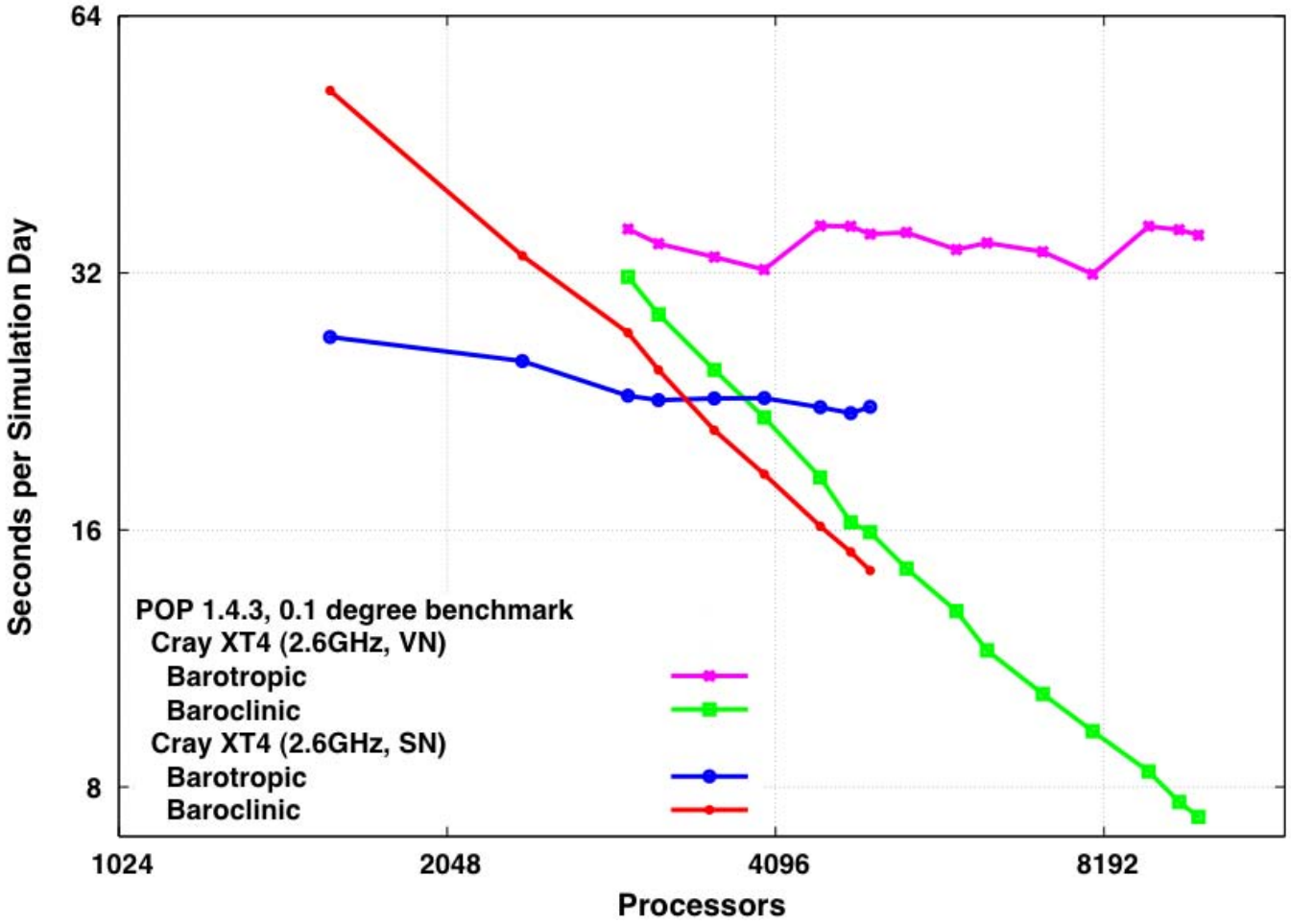


Images generated at NCAR using POP Explorer

POP Scaling

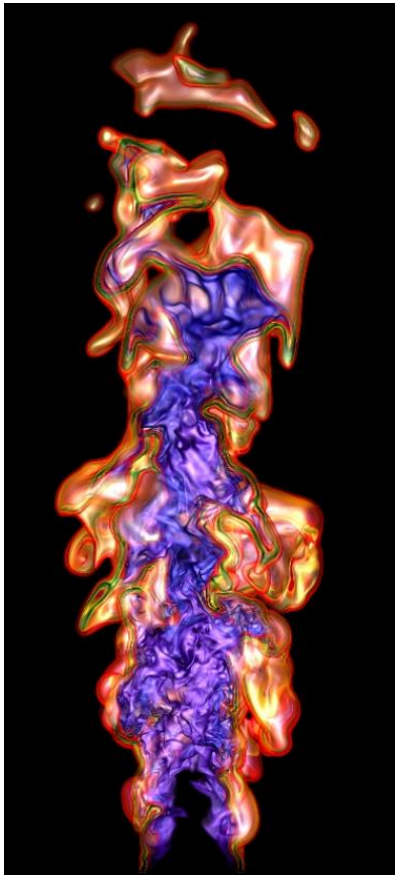

Simulation Years per Day
MPI Tasks
0
2
4
6
8
10
12
0
2000
4000
6000
8000
10000
POP 1.4.3, 0.1 degree benchmark
Cray XT4 (dual-core, 2.6GHz, SN)
Cray XT3 (dual-core, 2.6GHz, SN)
Cray XT4 (dual-core, 2.6GHz, VN)
Cray XT3 (dual-core, 2.6GHz, VN)
Cray XT3 (single-core, 2.4GHz)

Per Phase Performance and Scaling



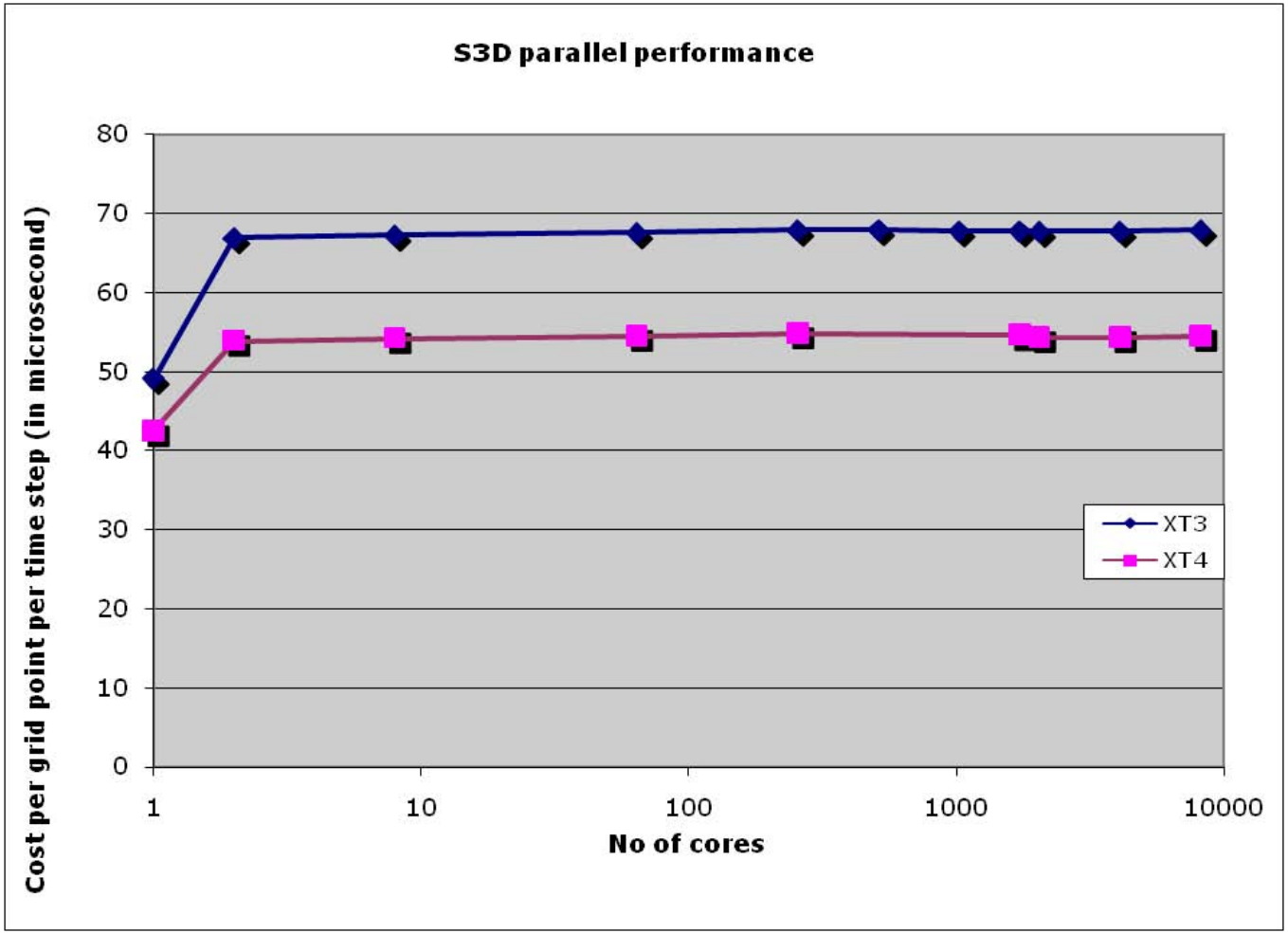
S3D - Combustion Solver

- Compressible flow solver for combustion science
- Detailed chemical kinetics and molecular transport using CHEMKIN library
- Explicit integration along with compressible formulation
- Mainly nearest neighbor communication



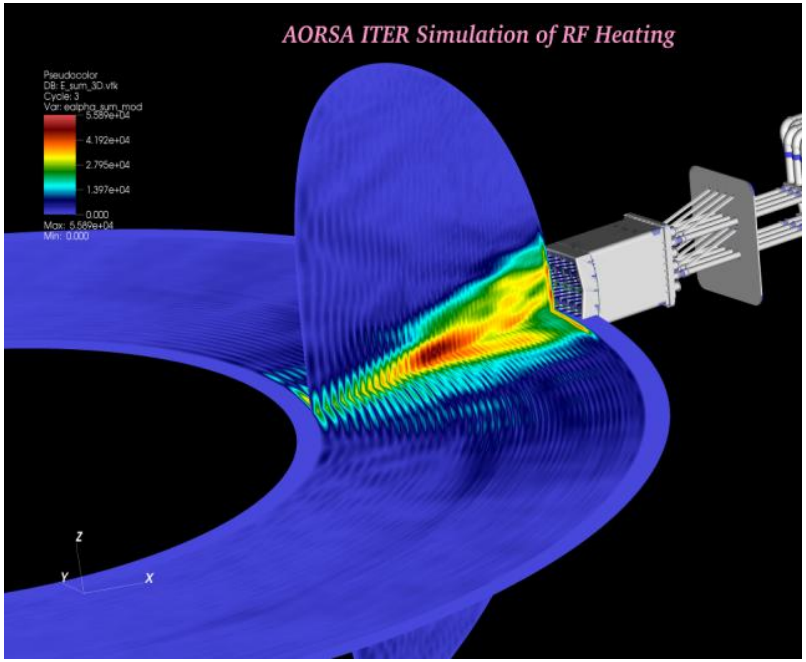
Turbulent premixed combustion simulated using S3D. Visualization by Yu and Ma, SciDAC Ultrascale viz. institute.

Direct numerical simulation (DNS) using S3D



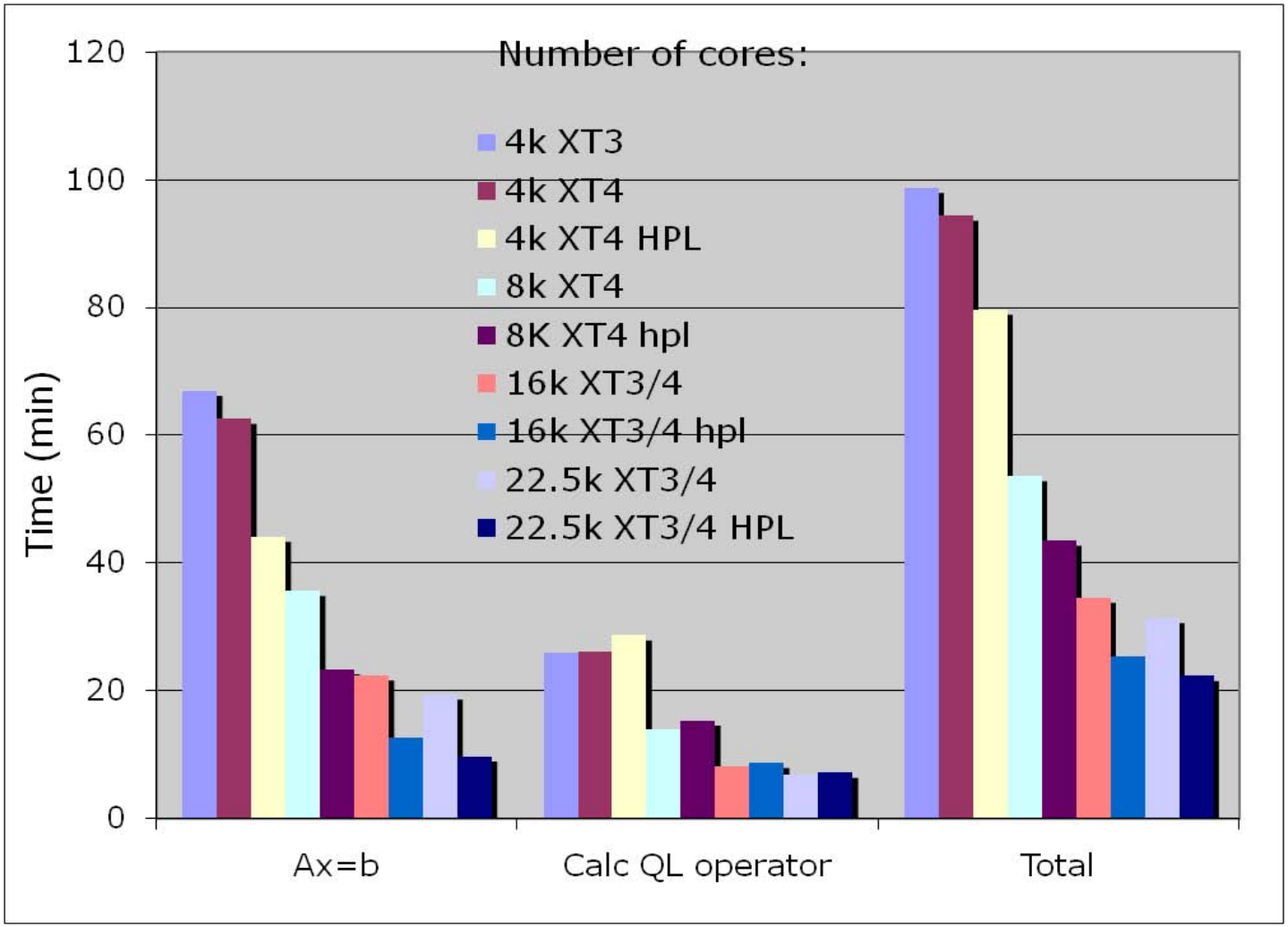
Fusion Energy

- All Order Spectral Algorithms (AORSA) application which models rf heating of plasma
  - Part of ITER project
- FFT and dense complex-coefficient system of linear equations
  - Exploits HPL (modified for complex-valued variables) and GotoBLAS



Sean Ahern and Fred Jaeger (ORNL)

All Order Spectral Algorithms (AORSA)



Conclusions and future plans

- A very controlled set of experiments reveal
  - System balance is critical
    - One parameter does not yield performance
  - Inherent scaling limits can offset all benefits
- To do items
  - Re-evaluate impact of improved MPI collectives
  - IO sensitivity studies
  - Advantages of additional programming models supported by Compute Node Linux (CNL)
  - Additional metrics for application characterization on multi-core MPP systems

OAK RIDGE National Laboratory

Questions