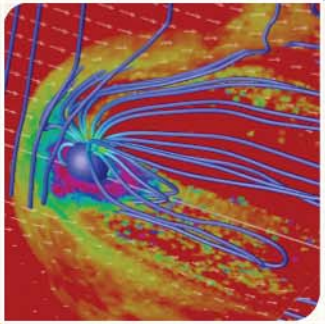
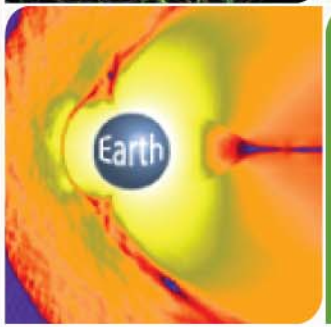
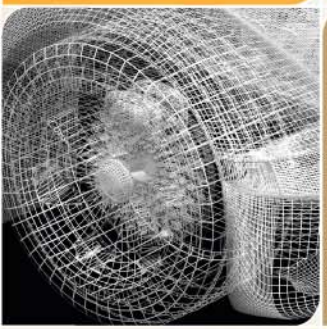



IAA IC Workshop & Apps Characterization

Jeffrey Vetter

Oak Ridge National Laboratory and Georgia Tech



Georgia Tech College of Computing

Computational Science and Engineering

http://www.csm.ornl.gov/workshops/IAA-IC-Workshop-08/

July 21-22, 2008

IAA Interconnection Network Workshop

San Jose, California

- Home
- Presentations
- Agenda (pdf)
- Attendees (pdf)
- Lodging
- Transportation
- Area Attractions

IAA Interconnection Network Workshop

Sandia and Oak Ridge National Laboratories have established the Institute for Advanced Architectures and Algorithms (IAA) to address critical issues in the design of next generation HPC architectures. Clearly, interconnection networks play a critical role in the architecture of peta- and exa-scale computing systems. As such, the IAA will bring together experts from academia, industry, and government at this workshop to identify critical challenges for next-generation interconnects. Topics will include algorithms, network architecture, device technology, programming models, architecture simulation, and system software.

This workshop is by invitation only.

This year's conference organizers are:

Co-chairs

Scott Hemmert - Sandia National Laboratories
Jeffrey Vetter - Oak Ridge National Laboratory and Georgia Tech

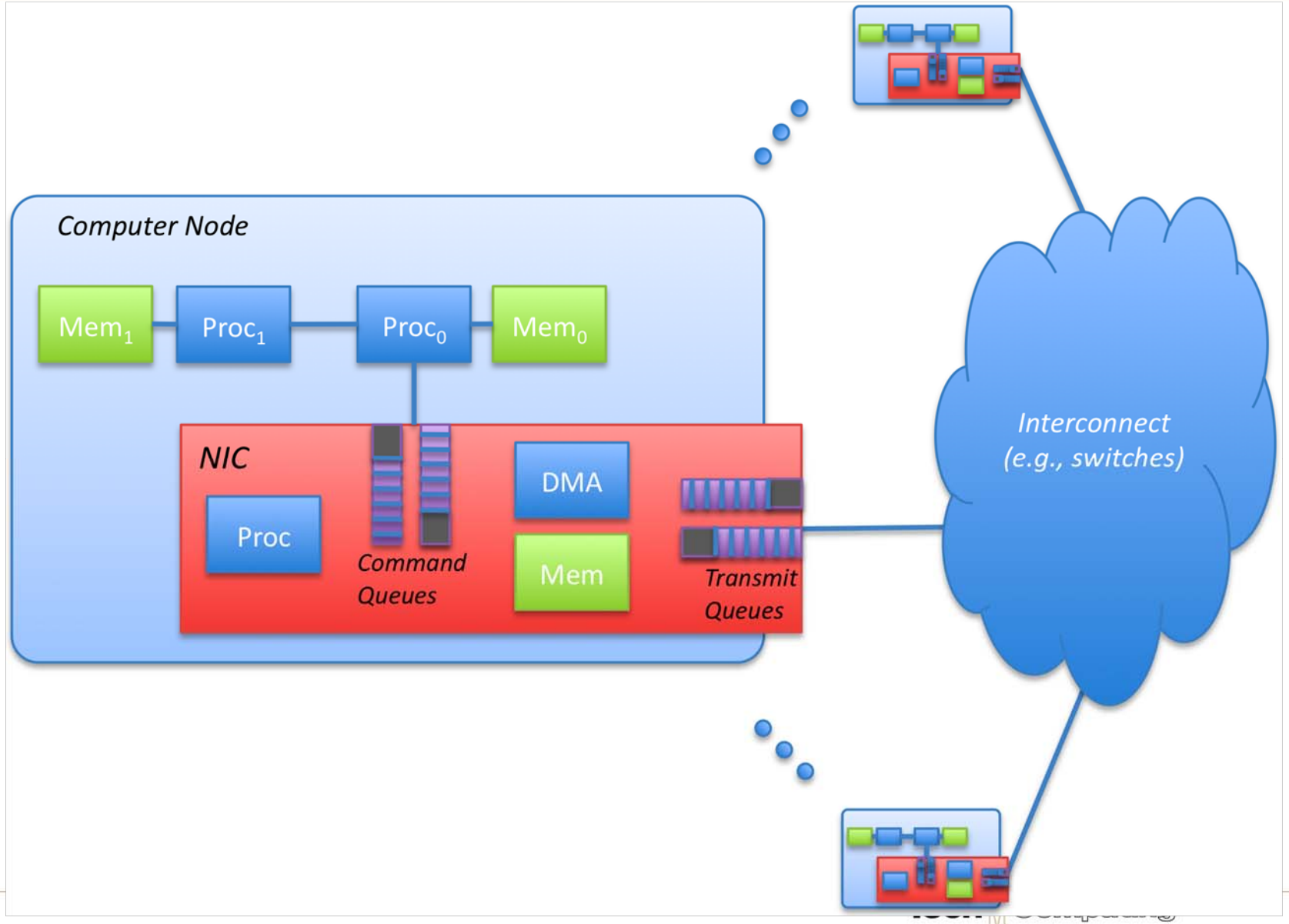
Organizing Committee

Keren Bergman - Columbia
Ron Brightwell - Sandia National Laboratories
Candice Culhane - DOD
Bill Dally - Stanford University
Bill Harrod - DARPA
Scott Hemmert - Sandia National Laboratories
Thuc Hoang - NNSA
Fred Johnson - DOE Office of Science
Scott Pakin - Los Alamos National Laboratories
Jeffrey Vetter - Oak Ridge National Laboratory and Georgia Tech

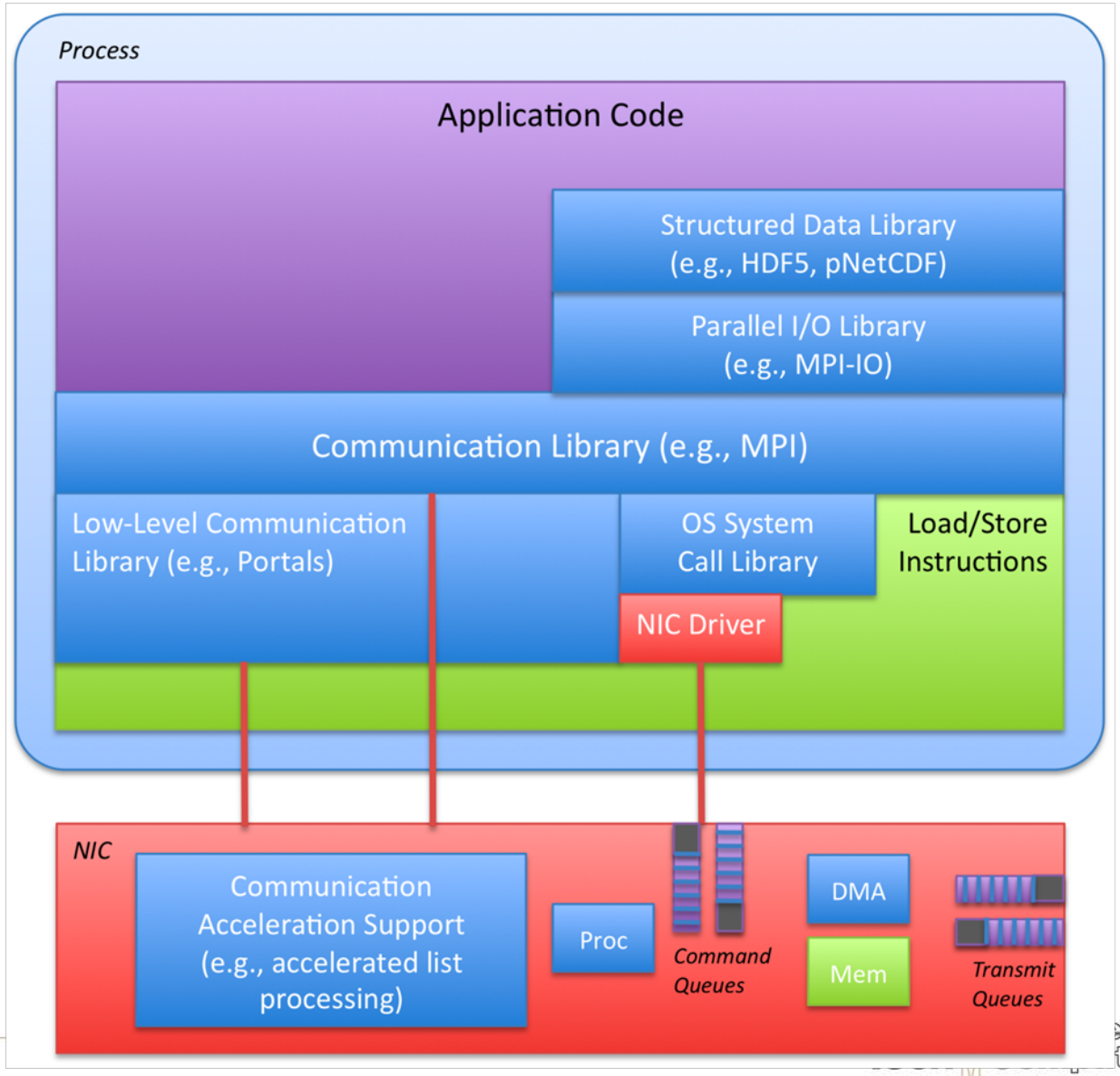
Goal – Identify IC Challenges for Exascale

- Motivation and Background
  - Exascale design points
  - Applications
  - Programming Systems
  - System software
- 76 Attendees
- Working groups
  - Topology and Routing
    - Craig Stunkel
    - Chita Das
  - Processor-Network interface
    - DK Panda
    - Keith Underwood
  - Simulation and Performance Prediction
    - Curtis Janssen
    - Sudhakar Yalamanchili
  - IC Technology
    - Keren Bergman
    - Azita Emami

Hardware


Computer Node
Mem1
Proc1
Proc0
Mem0
NIC
Proc
Command Queues
DMA
Mem
Transmit Queues
Interconnect (e.g., switches)

Software

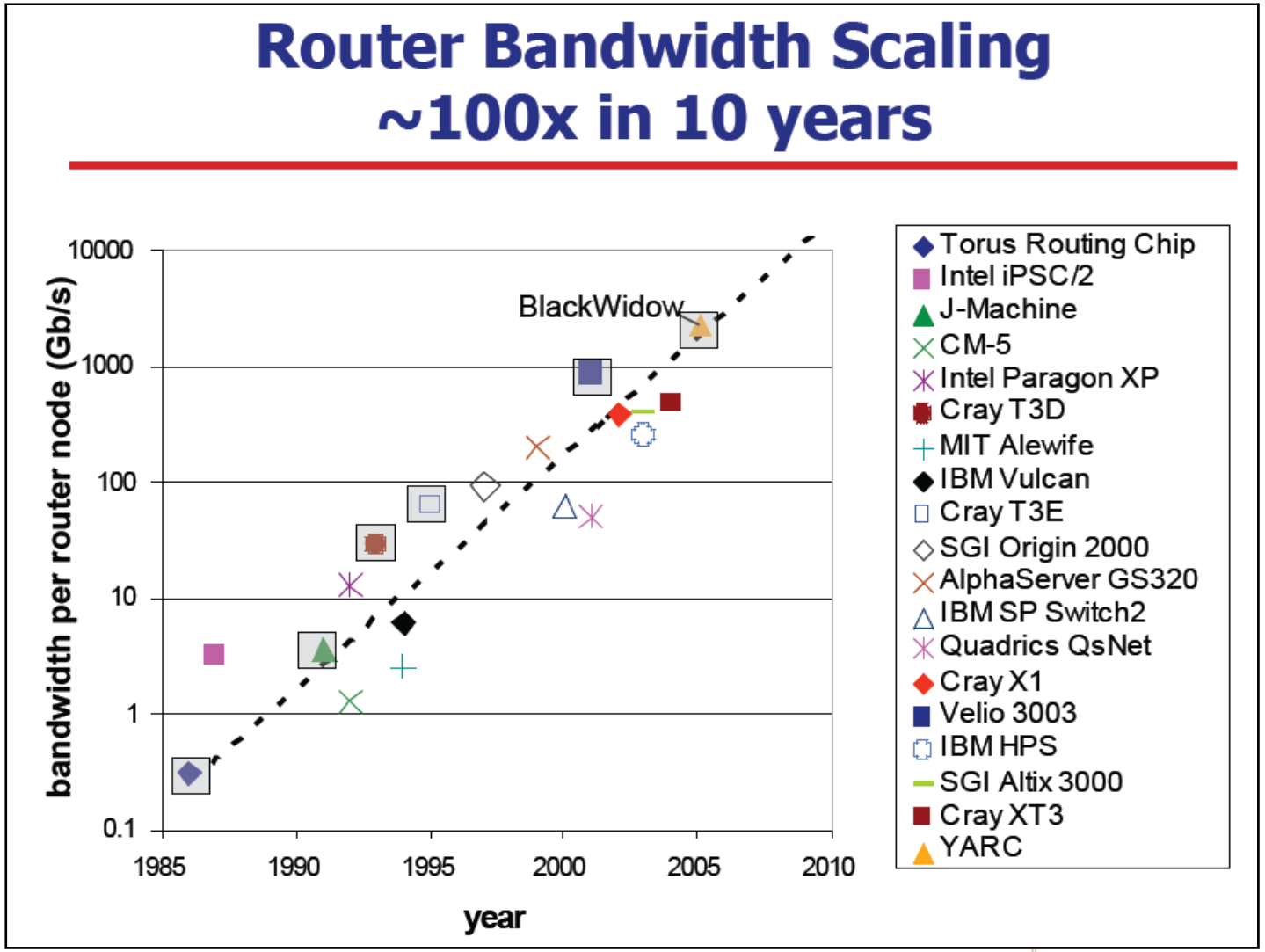

Process
Application Code
Structured Data Library
(e.g., HDF5, pNetCDF)
Parallel I/O Library
(e.g., MPI-IO)
Communication Library (e.g., MPI)
Low-Level Communication
Library (e.g., Portals)
OS System
Call Library
Load/Store
Instructions
NIC Driver
NIC
Communication
Acceleration Support
(e.g., accelerated list
processing)
Proc
Command
Queues
DMA
Mem
Transmit
Queues

Skeleton Configurations/Expectations

	2011		2014		2018	
System Size Sockets Peak PF TF/Socket	32,768 32 1.0		32,768 200 6.1		32,768 800 25.0	
	Expect	Want	Expect	Want	Expect	Want
NIC B/W (B/F)	0.01 - 0.1	1.0	0.005 - 0.03	1.0	0.025 - 0.25	1.0
Link B/W (B/F)	0.01 - 0.1	1.0	0.005 - 0.03	1.0	0.025 - 0.25	1.0
MPI Latency (ns)	750 - 1500	500	500 - 1000	400	400 - 750	300
MPI Throughput (M Msg/s)	20	50	80	300	300	1200
Load/Store (M Msg/s)	75	400	150	1,600	300	6400
Load/Store Latency (ns)	300	100	300	100	300	100

Source: IAA Report & J. Tomkins, "Interconnects: A Buyers Point of View," Sandia National Laboratories ACS613-1407, 2007.

Router Bandwidth Scaling



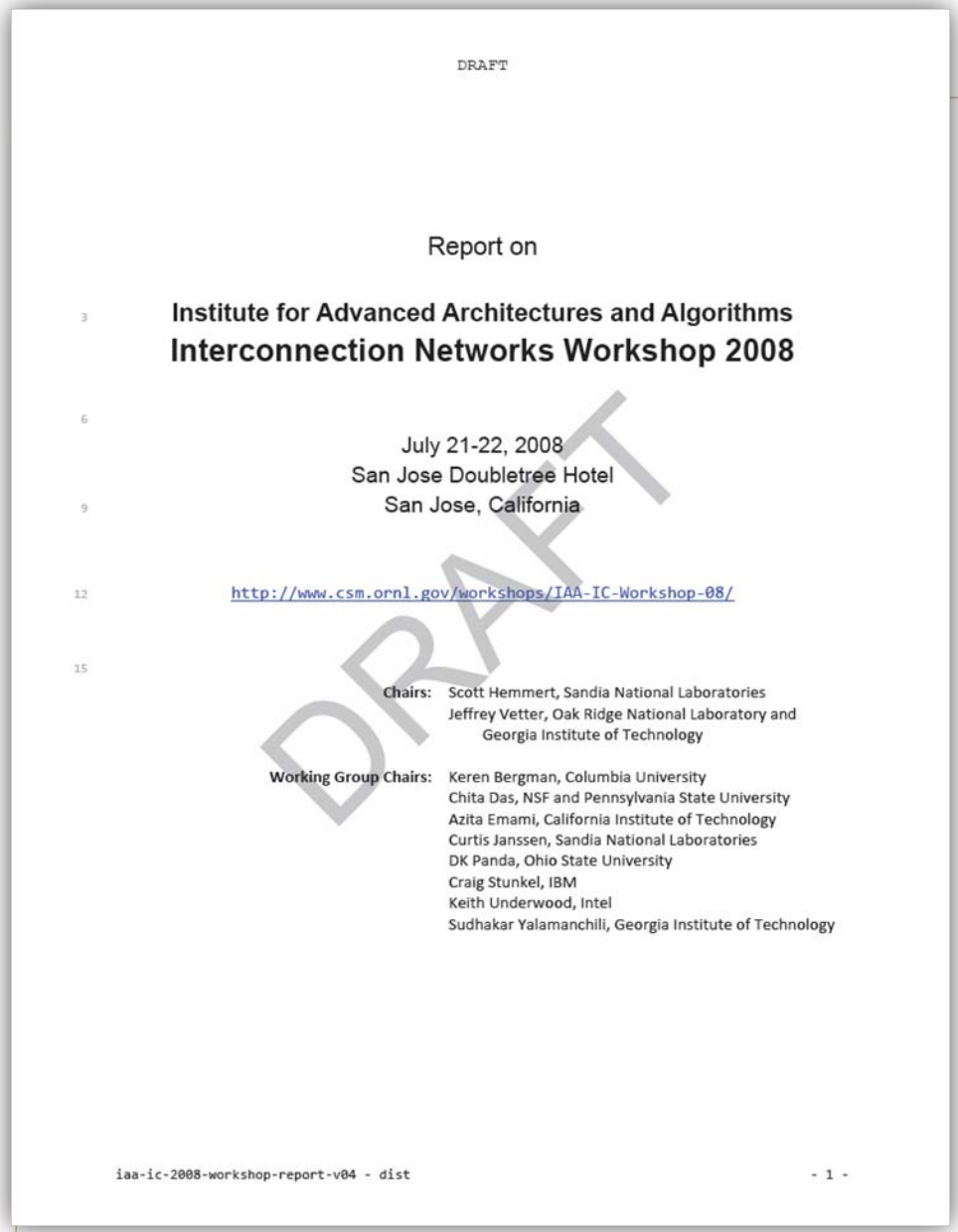
Source: Dally, Kim; IAA IC Workshop

Preliminary Findings (1)

- Topology and Routing
  - Minimize energy per bit
  - Maximize bandwidth per cost for both local and global patterns
  - Reliability
  - Optimal topology for integrated networks
  - Quality of Service and congestion control
  - Minimize latency for small messages
- Processor Network Interface
  - On-chip and Off-chip NI integration
  - Light-weight communication protocol
  - Enhanced NIC design
    - HW acceleration
    - Minimize cache pollution
  - End-to-End reliability
  - Collectives
  - Fine-grained synchronization
  - Connection management scalability
  - Converged network interface

Preliminary Findings (2)

- Simulation and Performance Prediction
  - Technology for composable simulators
  - Methodologies for constructing simulators
  - Model calibration
  - Simulator calibration
  - Parallelism
  - Multi-scale
  - Hardware acceleration
  - Power and thermal models
- Device technology
  - Energy per bit
  - Clear leaders
    - CMOS Photonics
    - Proximity communication
    - MEMS
    - 3-D Stacking
  - Uncertain
    - Carbon nanotubes
    - Free-space optics
    - RF communications



DRAFT

Report on

Institute for Advanced Architectures and Algorithms
Interconnection Networks Workshop 2008

July 21-22, 2008
San Jose Doubletree Hotel
San Jose, California

http://www.csm.ornl.gov/workshops/IAA-IC-Workshop-08/

Chairs: Scott Hemmert, Sandia National Laboratories
Jeffrey Vetter, Oak Ridge National Laboratory and Georgia Institute of Technology

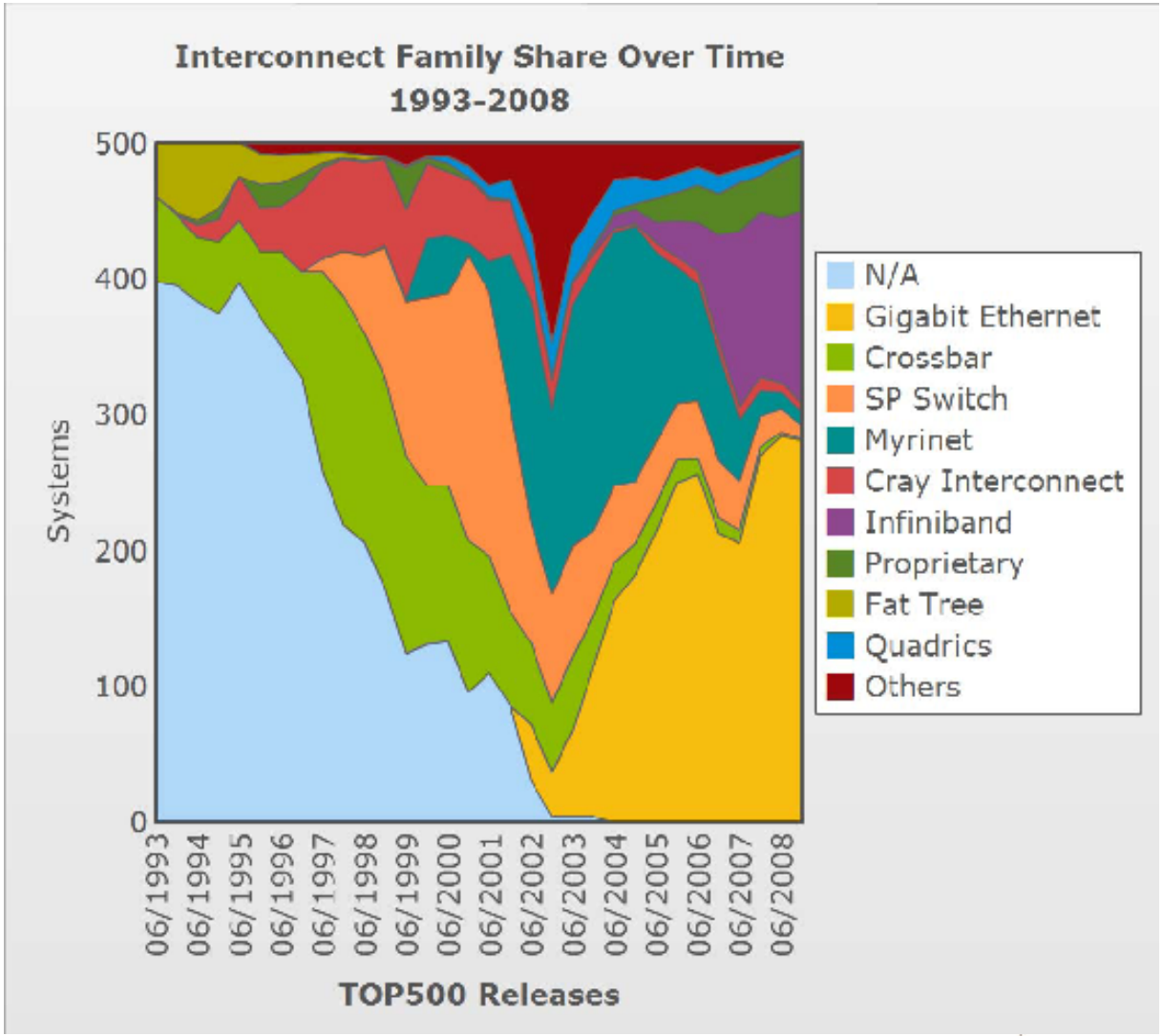
Working Group Chairs: Keren Bergman, Columbia University
Chita Das, NSF and Pennsylvania State University
Azita Emami, California Institute of Technology
Curtis Janssen, Sandia National Laboratories
DK Panda, Ohio State University
Craig Stunkel, IBM
Keith Underwood, Intel
Sudhakar Yalamanchili, Georgia Institute of Technology

DRAFT

iaa-ic-2008-workshop-report-v04 - dist - 1 -

- Report is being completed this month
- To be published on the workshop website

Commodity Interconnects dominate TOP500



Source: TOP500.org

Commodity represents 40% of TOP 10!

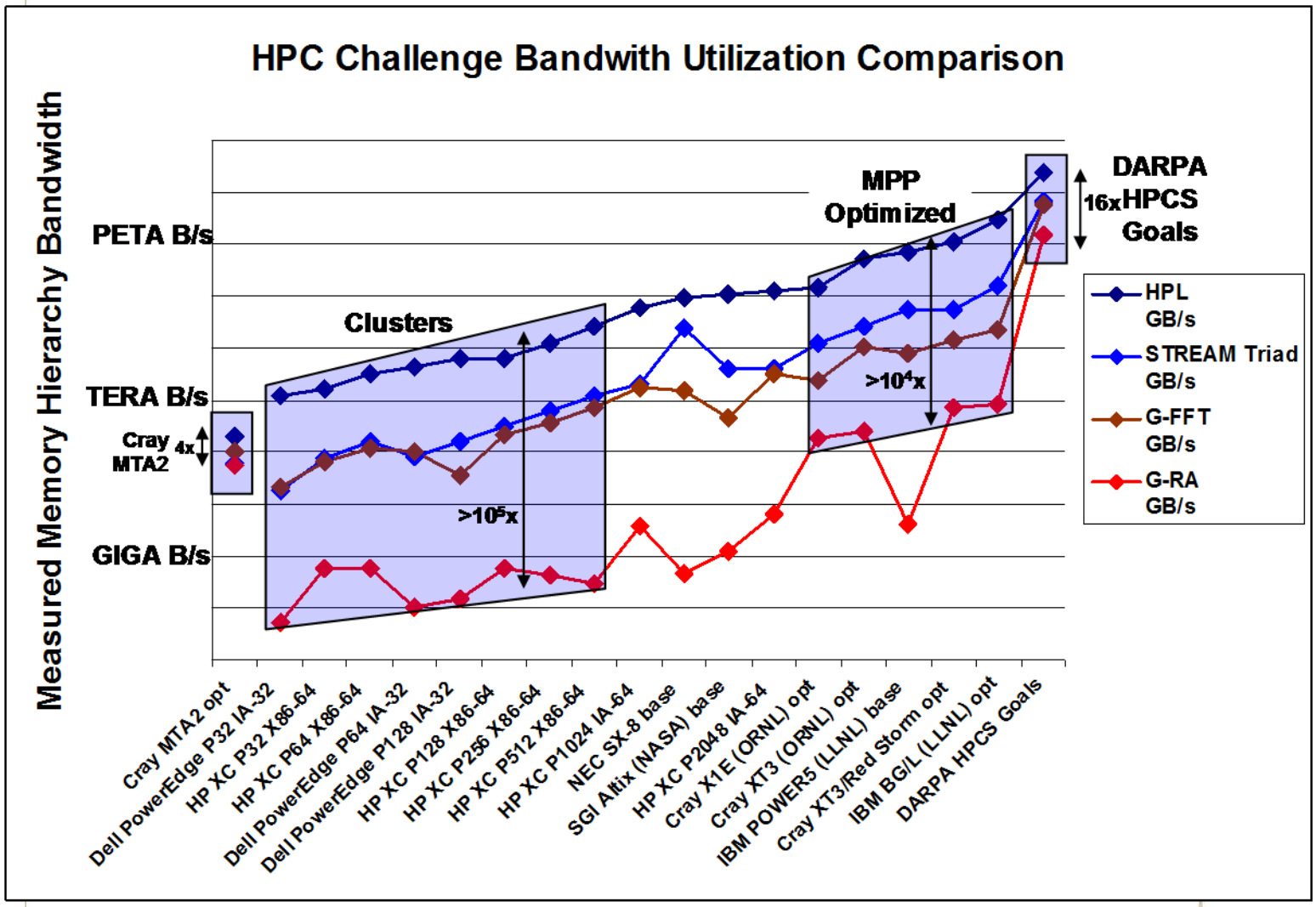
- TOP 10 Systems
  - IB 40%
  - Cray XT Torus: 40%
  - IBM BlueGene Torus/Tree: 20%

Rank	Site	Manufacturer	Computer	System Model	Interconnect Family	Interconnect
1	DOE/NNSA/LANL	IBM	BladeCenter QS22/LS21 Cluster, F	BladeCenter QS22 Cluster	Infiniband	Infiniband
2	Oak Ridge National Laboratory	Cray Inc.	Cray XT5 QC 2.3 GHz	Cray XT5 QuadCore	Proprietary	XT4 Internal Interconnec
3	NASA/Ames Research Center/NAS	SGI	SGI Altix ICE 8200EX, Xeon QC 3.	SGI Altix ICE 8200	Infiniband	Infiniband
4	DOE/NNSA/LLNL	IBM	eServer Blue Gene Solution	BlueGene/L	Proprietary	Proprietary
5	Argonne National Laboratory	IBM	Blue Gene/P Solution	BlueGene/P	Proprietary	Proprietary
6	Texas Advanced Computing Center/Univ	Sun Microsyst	SunBlade x6420, Opteron QC 2.3	Sun Blade x6420	Infiniband	Infiniband
7	NERSC/LBNL	Cray Inc.	Cray XT4 QuadCore 2.3 GHz	Cray XT4 QuadCore	Proprietary	XT4 Internal Interconnec
8	Oak Ridge National Laboratory	Cray Inc.	Cray XT4 QuadCore 2.1 GHz	Cray XT4 QuadCore	Proprietary	XT4 Internal Interconnec
9	NNSA/Sandia National Laboratories	Cray Inc.	Sandia/ Cray Red Storm, XT3/4, 2.	Cray XT3	Cray Interconnec	XT3 Internal Interconnec
10	Shanghai Supercomputer Center	Dawning	Dawning 5000A, QC Opteron 1.9 G	Dawning 5000A	Infiniband	Infiniband DDR

Georgia Tech College of Computing

However, TOP500 & HPL are misleading

Select Systems from the 3 January 2007 HPC Challenge Benchmark Results



All benchmarks have been normalized as *Bandwidth Utilization* within the memory hierarchy

Data displayed are B/s = (B/op x op/s)

HPCS Goals *flatten* the memory hierarchy

HPCS Goals are similar to Cray MTA2 but at 10^5x the scale

Optimized G-RandomAccess results offer both *scalable* and *improved* performance (10-10^2x)

Source: DARPA HPCS, D. Koester

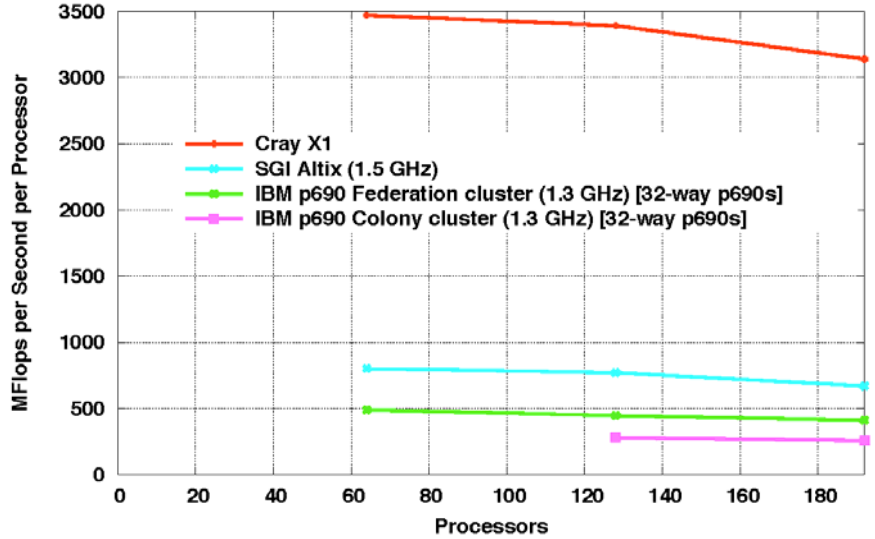
Real Applications have Wide Range of Characteristics

	Motifs											
Application	Structured Grid	Unstructured Grid	Multiresolution Grid	Dense Matrix	Sparse Matrix	Spectral/FFT	MD/N-Body	MapReduce / Monte Carlo	Graphs	PIC	Algo Load Imbalance	Adaptive
AMG					3							
IRS	3			3								
LAMMPS						3	3					
Sphot								3				
UMT		3										
GTC	2	2			3					3		
MILC	3				3	3				3		
DCA++				3				3				
GTC	2	2			3					3		
POP	3				3					3		
S3D	3											
SSCA2 V2.2									3			
GAMESS				3							1	
HYCOM	3											
WRF		3	3									1
FLASH												3
NAMD												
QBOX				3		3				3		
IOR												
PARATEC	3			3								
LSMS	3			3								

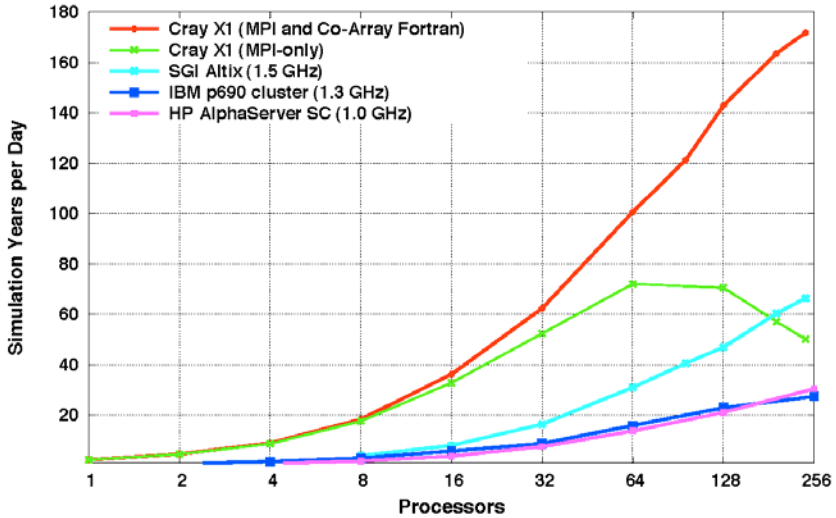
Georgia Tech College of Computing

Application Performance Limited by Interconnect Performance

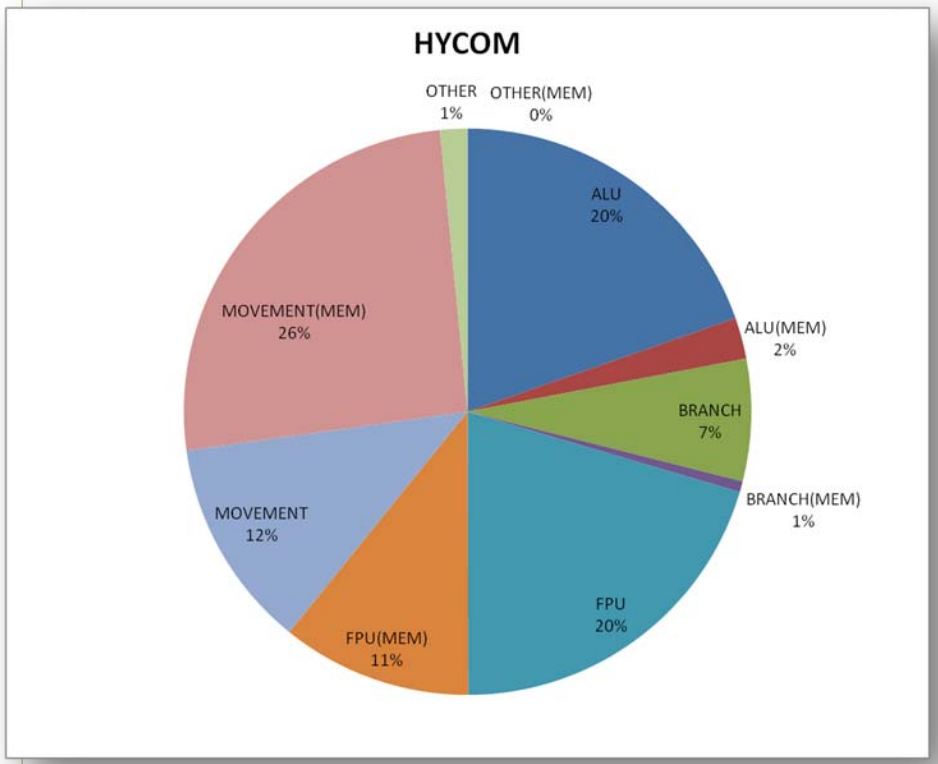
- Bandwidth limits GYRO performance
  - Significant increase in bandwidth dramatically improved GYRO performance

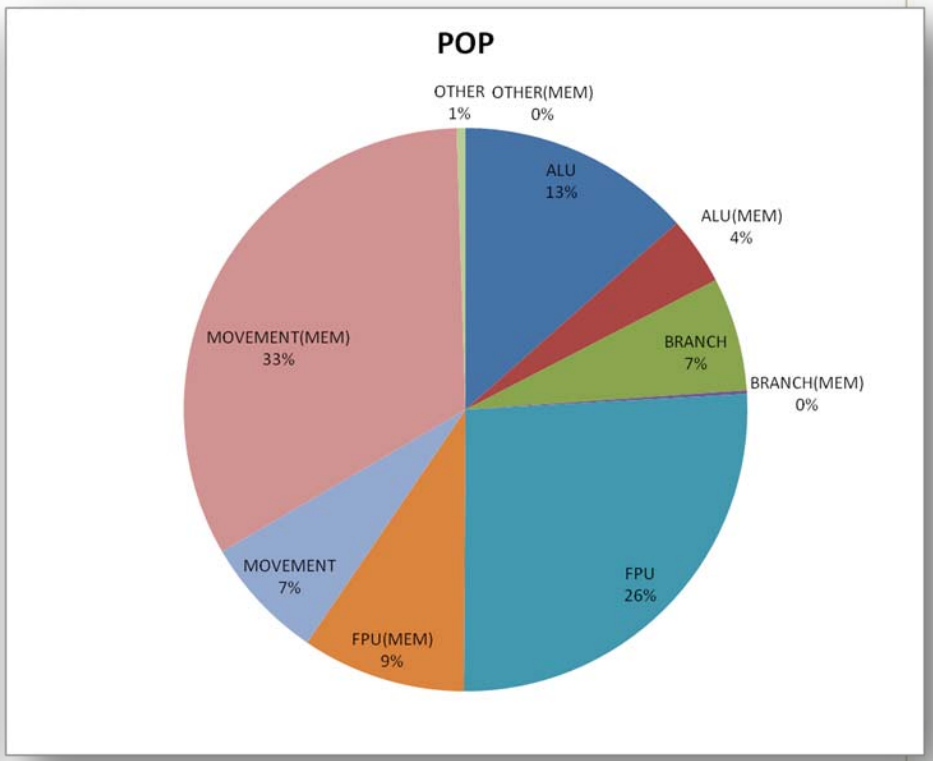


- Latency limits POP performance
  - Strong scaling problem
  - CAF mapping exploited X1 Globally addressable memory to provide dramatic scaling improvements



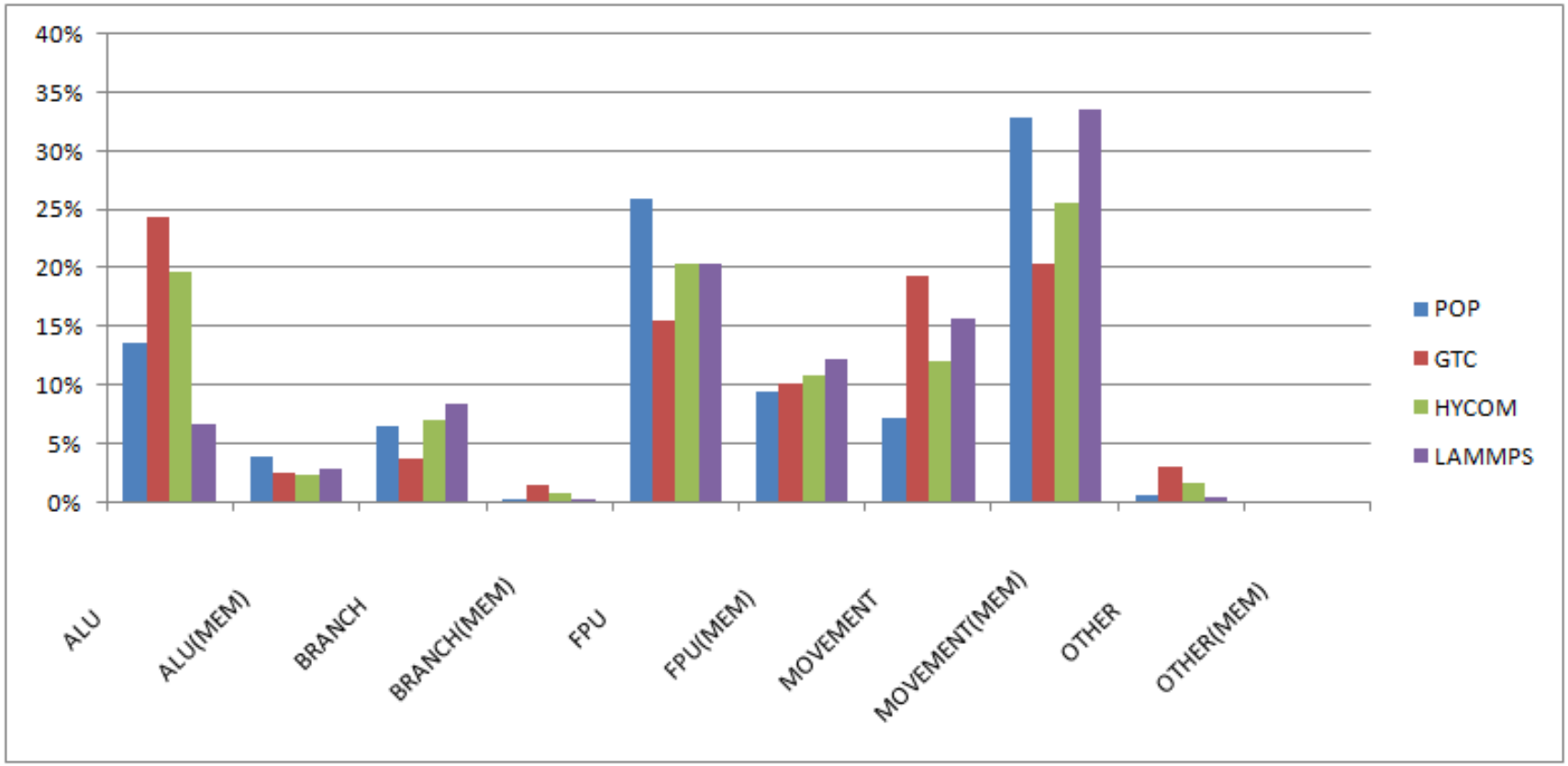
Only 15-26% of our applications instructions are FP





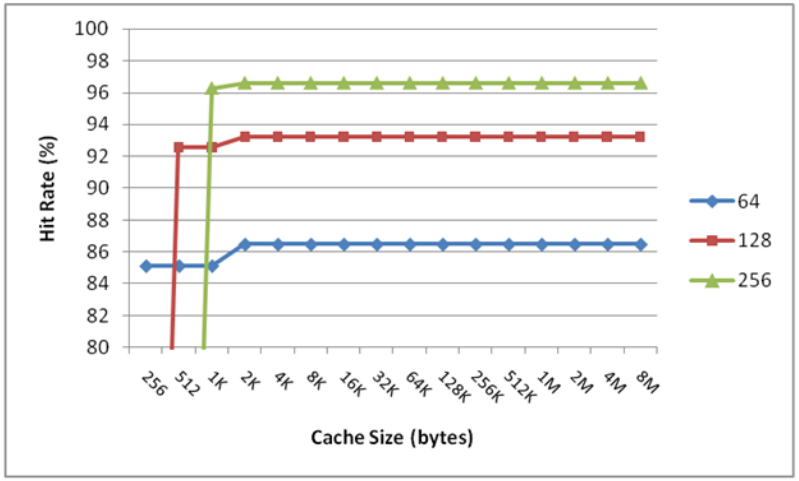
Thanks to Bermuda team for helping collect this data.
Phil Roth, Collin McCurdy.

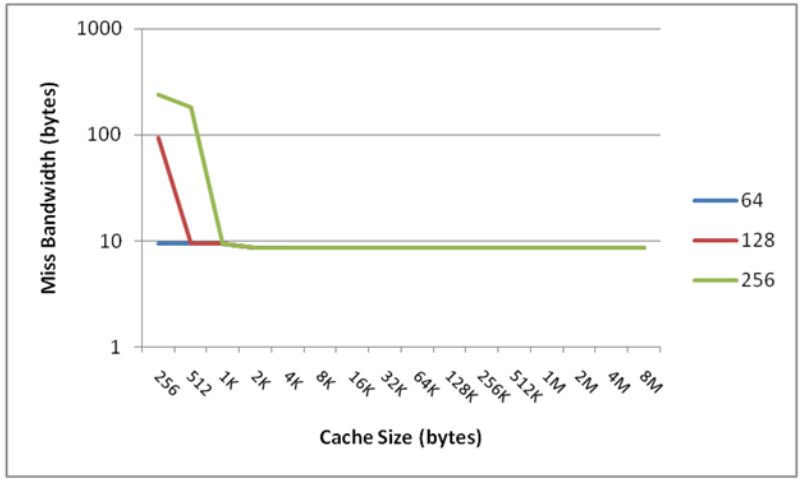
Data movement, ALU instructions dominate



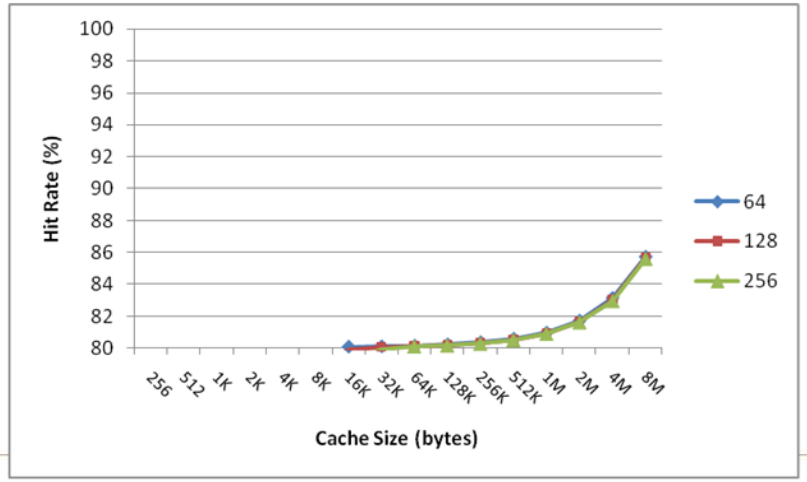
Locality will become more important as disparity increases

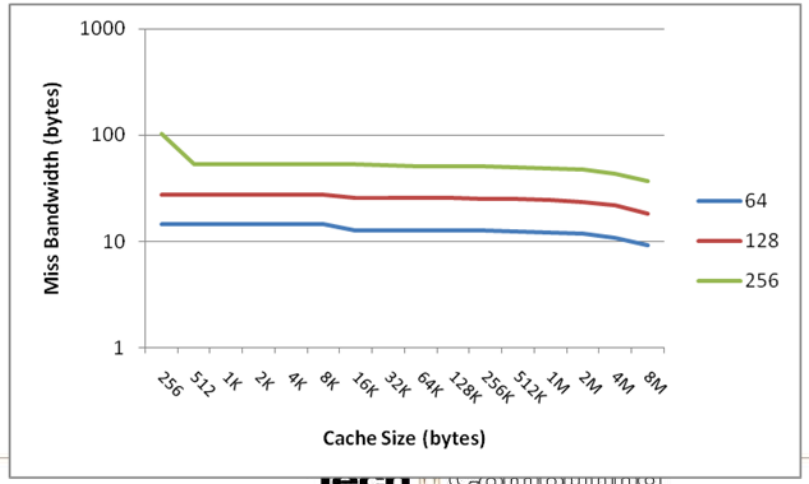
Stream





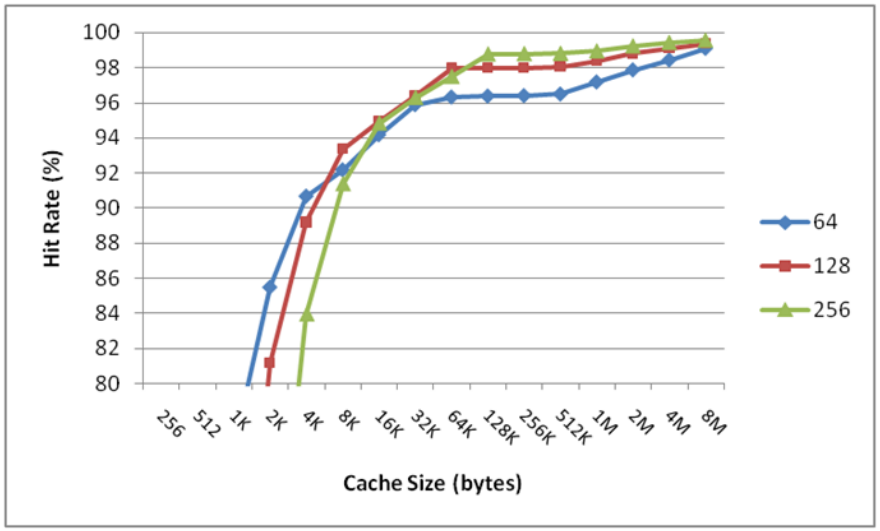
RA

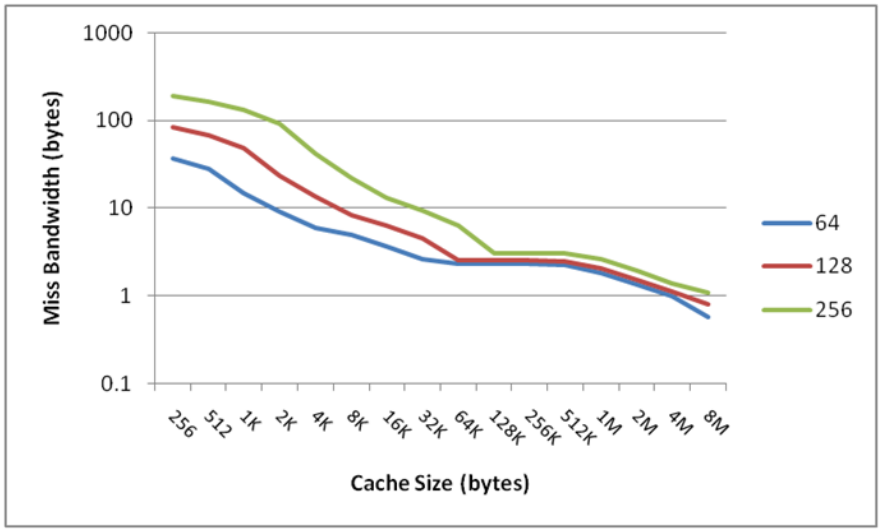




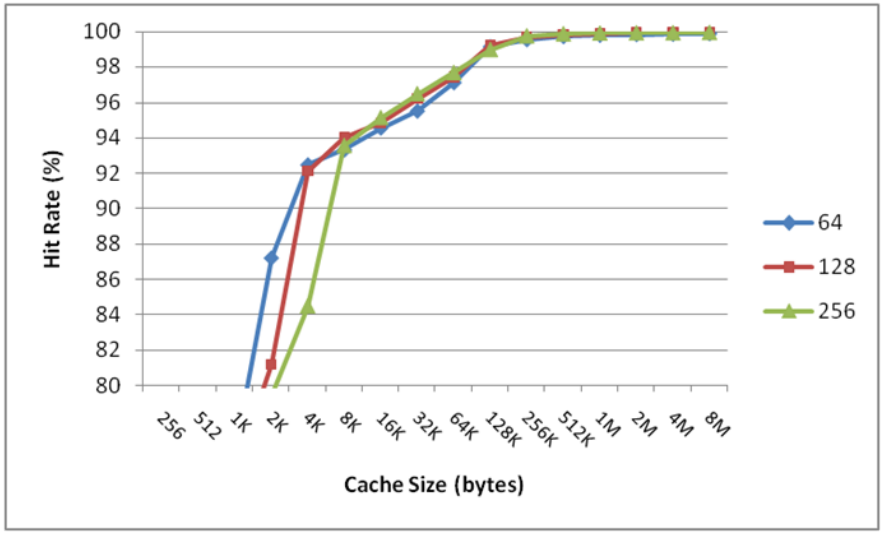
Locality will become more important as disparity increases (apps)

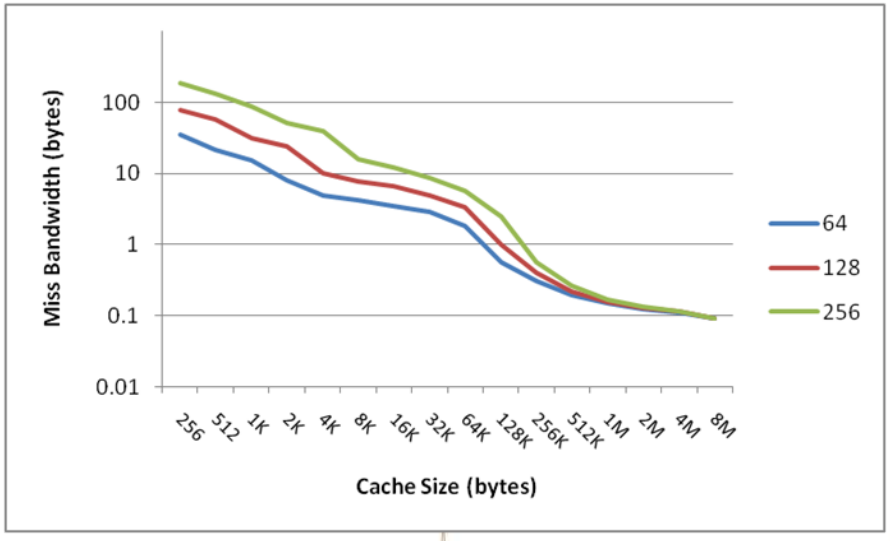
HYCOM





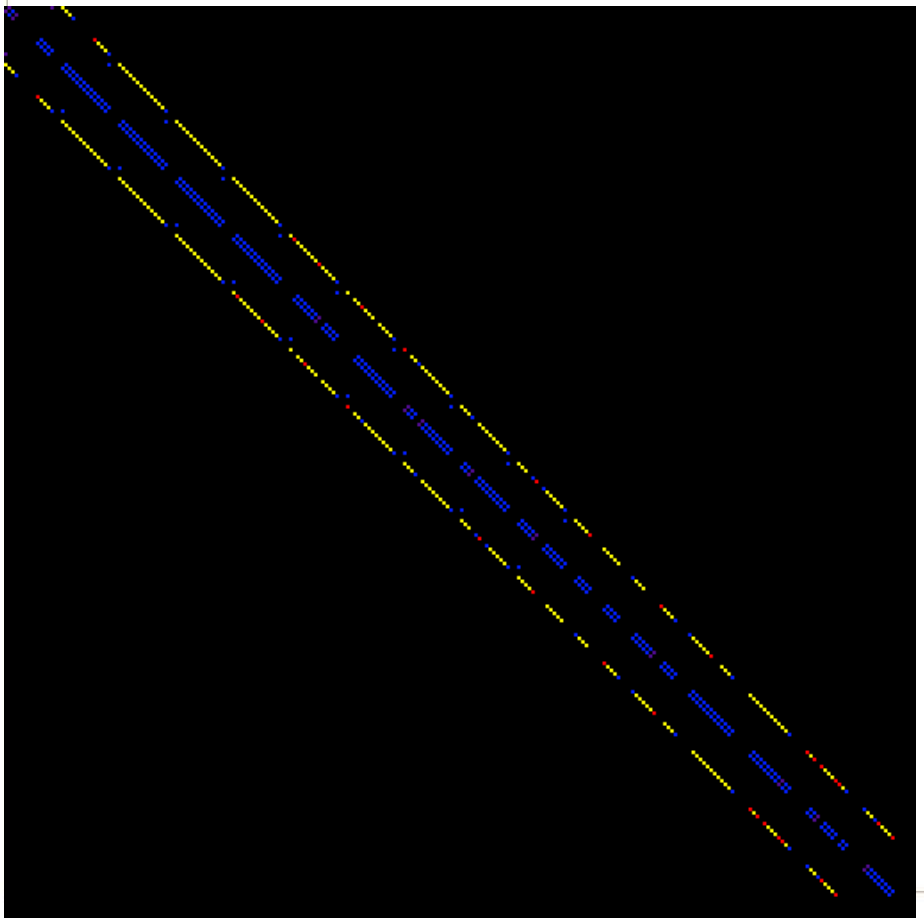
LAMMPS

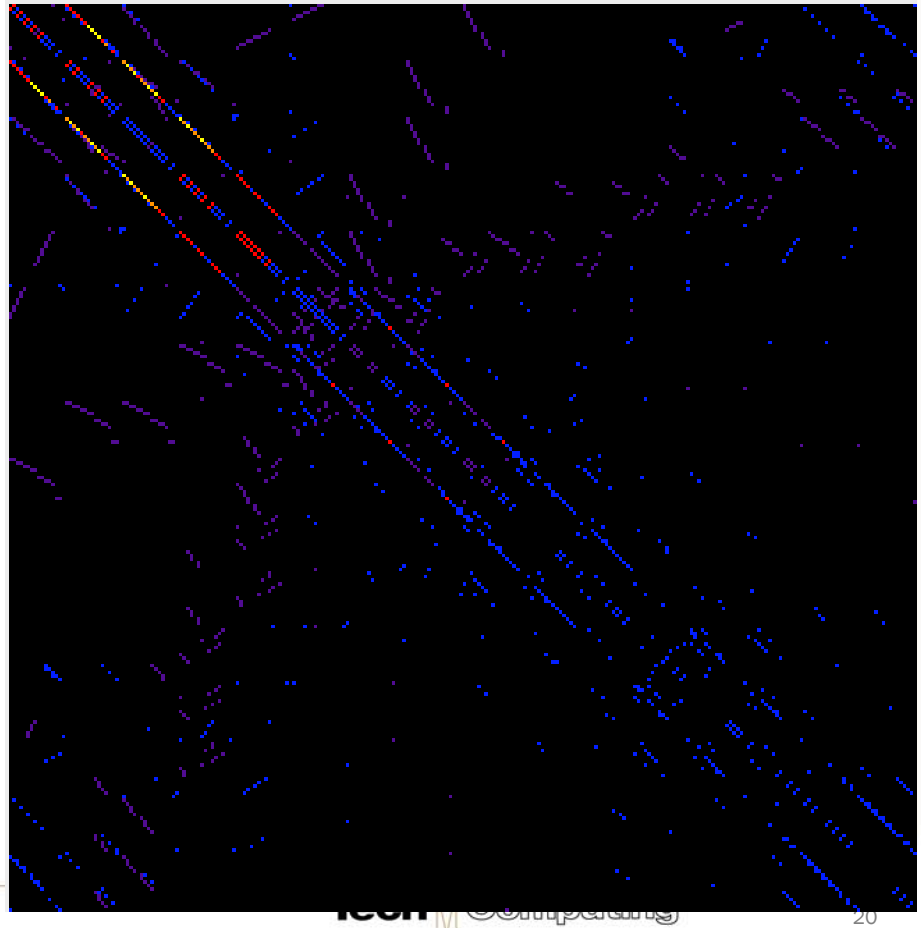




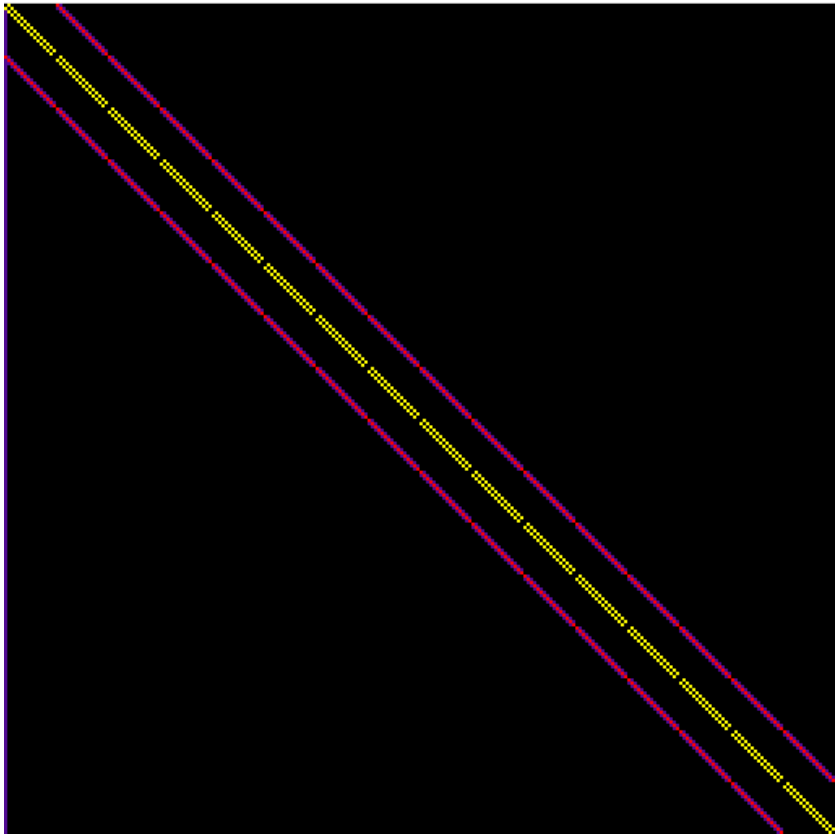
Logical Communication Topology: POP

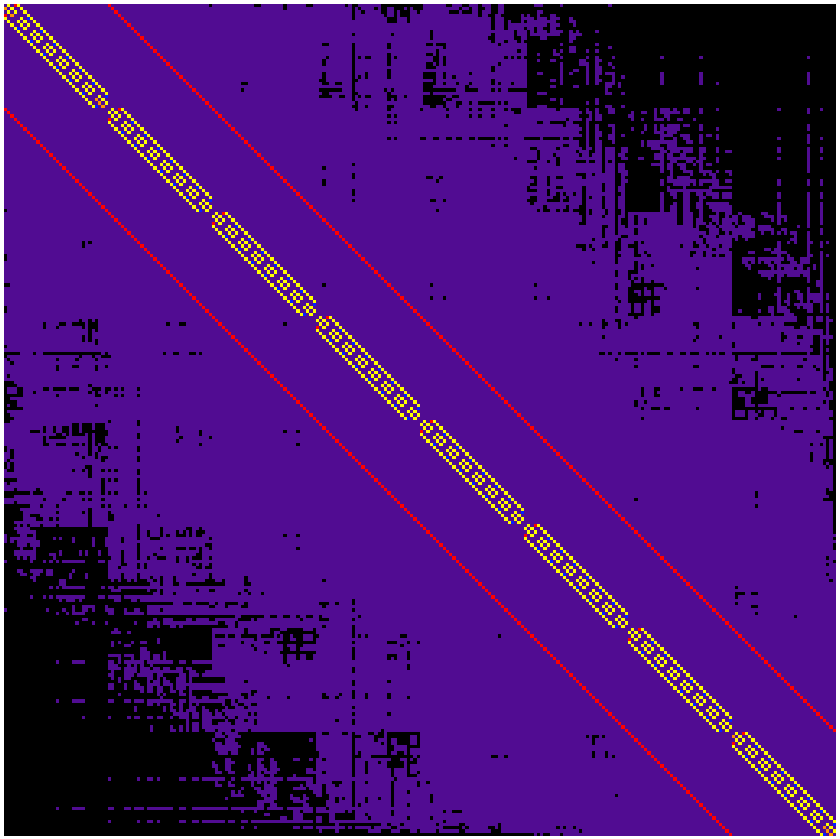
- POP regular
- POP with Load Balancing





Logical Communication Topology: WRF and Multigrid Solver





Application characterizations continue

- Work in Progress
  - Providing experimental confirmation of motifs
  - Continuing analysis of other real applications with simulation
- Are our applications more like HPL or RandomAccess?
  - Moving forward, it looks more like RA…
- Enhancements to programming models to
  - support fine-grained, one-sided messaging
  - Distributed transactional memory
  - Intelligent thread mapping

Summary

- IAA IC Workshop demonstrates interest from government, academia, industry
  - 76 attendees
  - Topology, simulation, processor network interface, device technology
  - Exascale within our constraints will not be possible on our current trajectory!
  - Some changes are not that expensive (photonic v. nic msg lists)
- Community must focus on 'real' workloads and kernels rather than Peak and HPL
  - How do we kill HPL?
    - HPL will require ~ 2 weeks to run on a 5 PF system!!
  - Use non-HPL benchmarks in design and procurements
    - HPCC? SPEC MPI 2007?
  - New benchmarks
- We are gaining a much better understanding of our applications using simulation and analysis

Acknowledgements and More Info

- Thanks to contributors, sponsors
  - Many collaborators across apps teams, academia, labs, and industry
  - DOE, ORNL, DOD, DARPA
- More information
  - http://ft.ornl.gov
  - vetter@computer.org
  - Publications: http://ft.ornl.gov/pubs

- Staff
  - Jeffrey Vetter (Group Leader)
  - Sadaf Alam
  - David Bader (joint w/ Georgia Institute of Technology)
  - Micah Beck (joint w/ University of Tennessee-Knoxville)
  - Anthony Danalis (joint w/ University of Tennessee-Knoxville)
  - Deb Holder (Group Secretary)
  - Gabriel Marin (in April)
  - Collin McCurdy
  - Jeremy Meredith
  - Kenneth Roche
  - Philip Roth
  - Thomas Sterling (joint w/ Louisiana State University)
  - Olaf Storaasli
  - Vinod Tipparaju
  - Weikuan Yu (joint w/ Auburn University)
  - Visitors
    - Ernest Cochran

This research was sponsored by the Office of Mathematical, Information, and Computational Sciences, Office of Science, U.S. Department of Energy under Contract No. DE-AC05-00OR22725 with UT-Battelle, LLC. Accordingly, the U.S. Government retains a non-exclusive, royalty-free license to publish or reproduce the published form of this contribution, or allow others to do so, for U.S. Government purposes.

Georgia Tech College of Computing