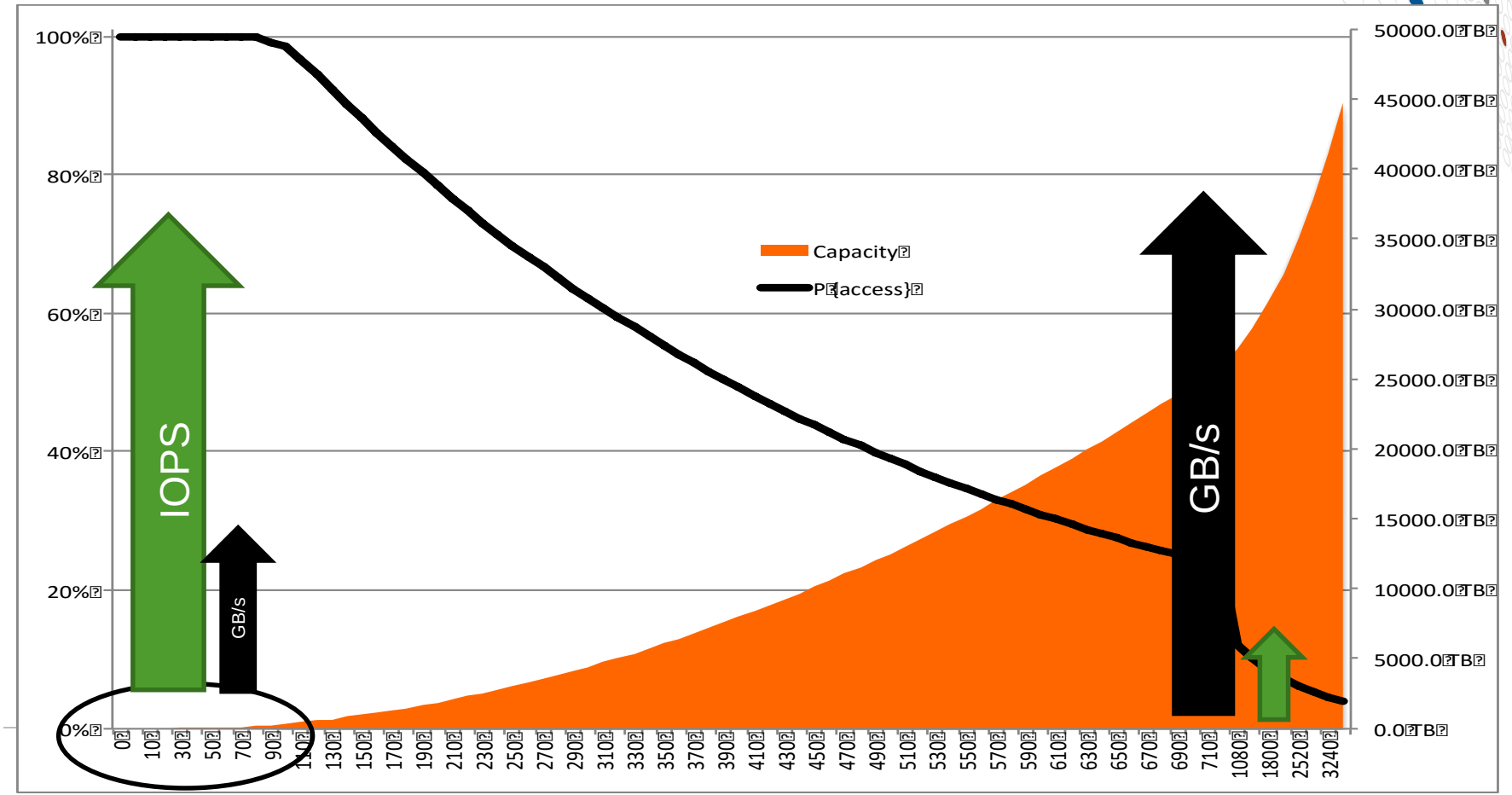
Sharing NVRAM and DRAM For Performance, Productivity, and Model Fidelity

June 2016

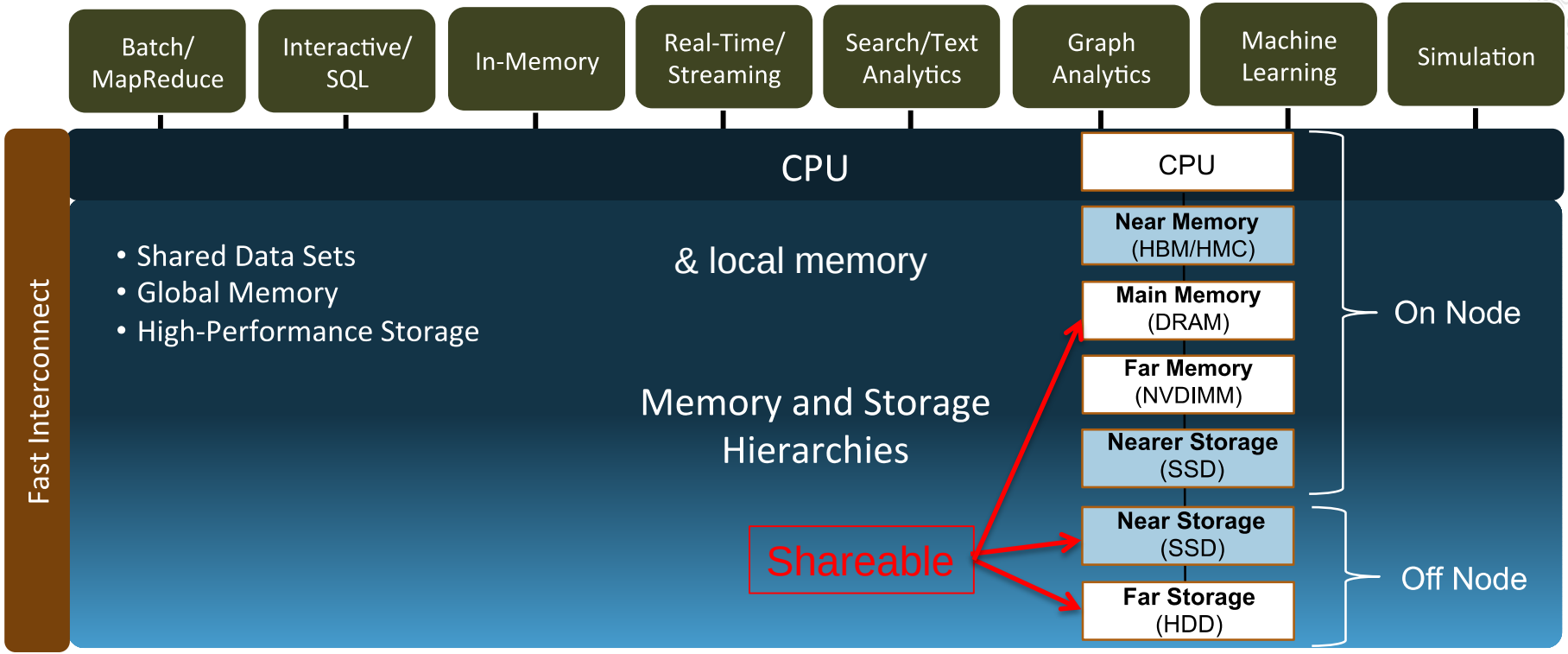
Philip Filleul – Segment Director FS

Differentiating Capacity and Bandwidth

REQUIREMENTS



Move up the Hierarchy for Performance
Share Data for Fidelity and Productivity



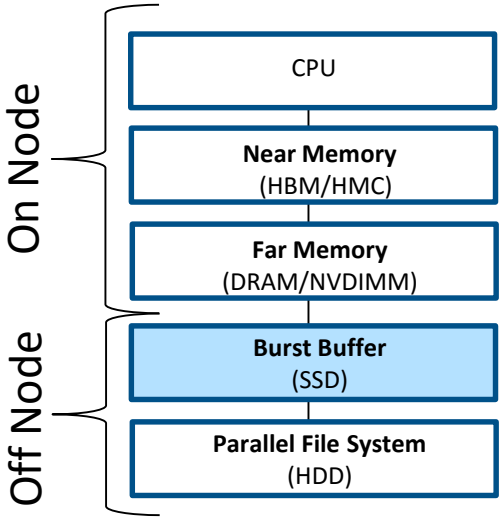
Cray Perspective on I/O in HPC

- **Expensive compute resources sit idle during I/O**
  - Want to have highest possible bandwidth when doing I/O
- **Disk-based PFS bandwidth is expensive**
  - Bandwidth via controllers inflates the effective cost
  - PFS is still the preferred option for scalability & permanence

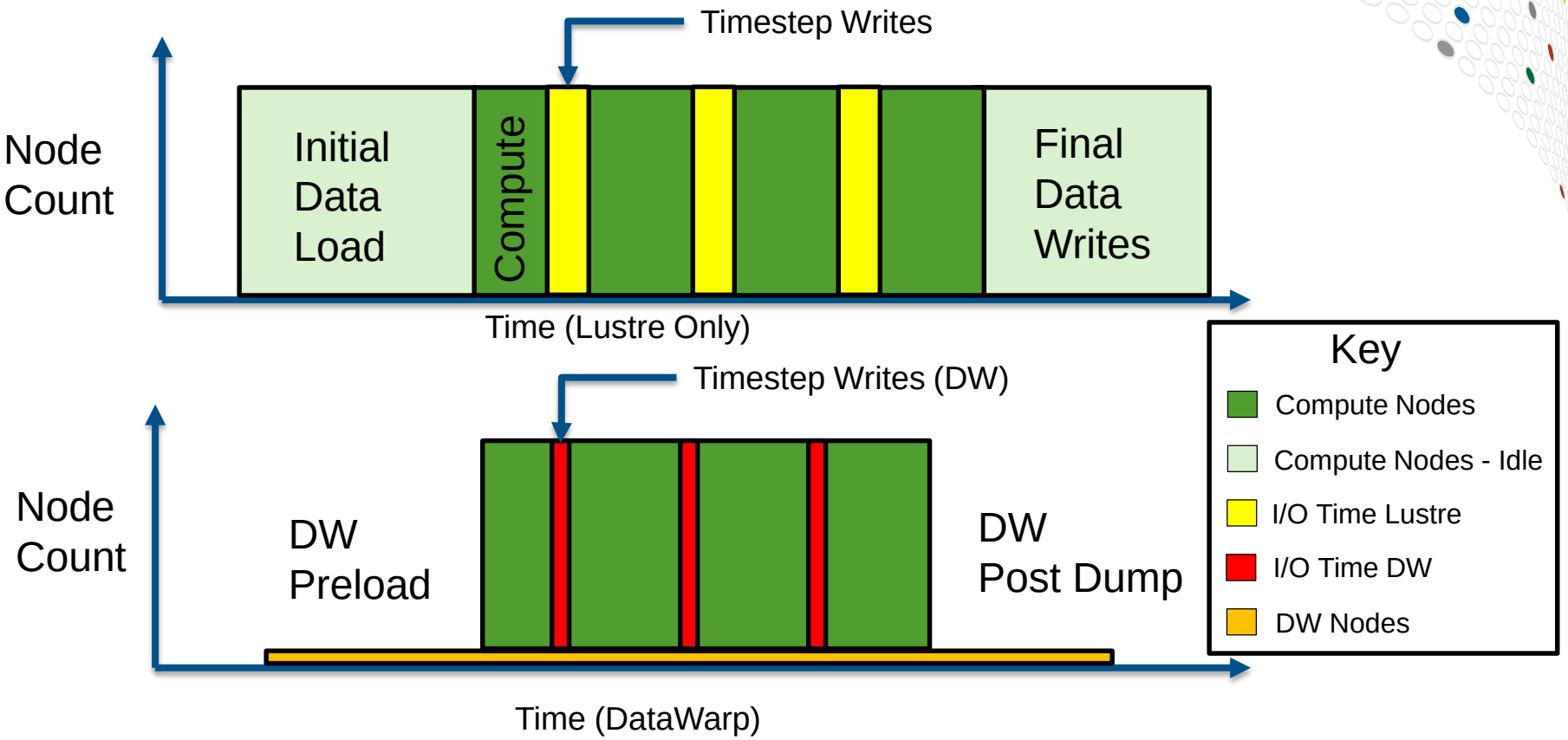
- **Many applications do I/O in bursts**
  - A cycle of Read → Compute → Write
  - Checkpoints
- **Flash bandwidth is relatively inexpensive**
  - Effective for I/O load at beginning/end of job, during checkpointing

The Burst Buffer Concept

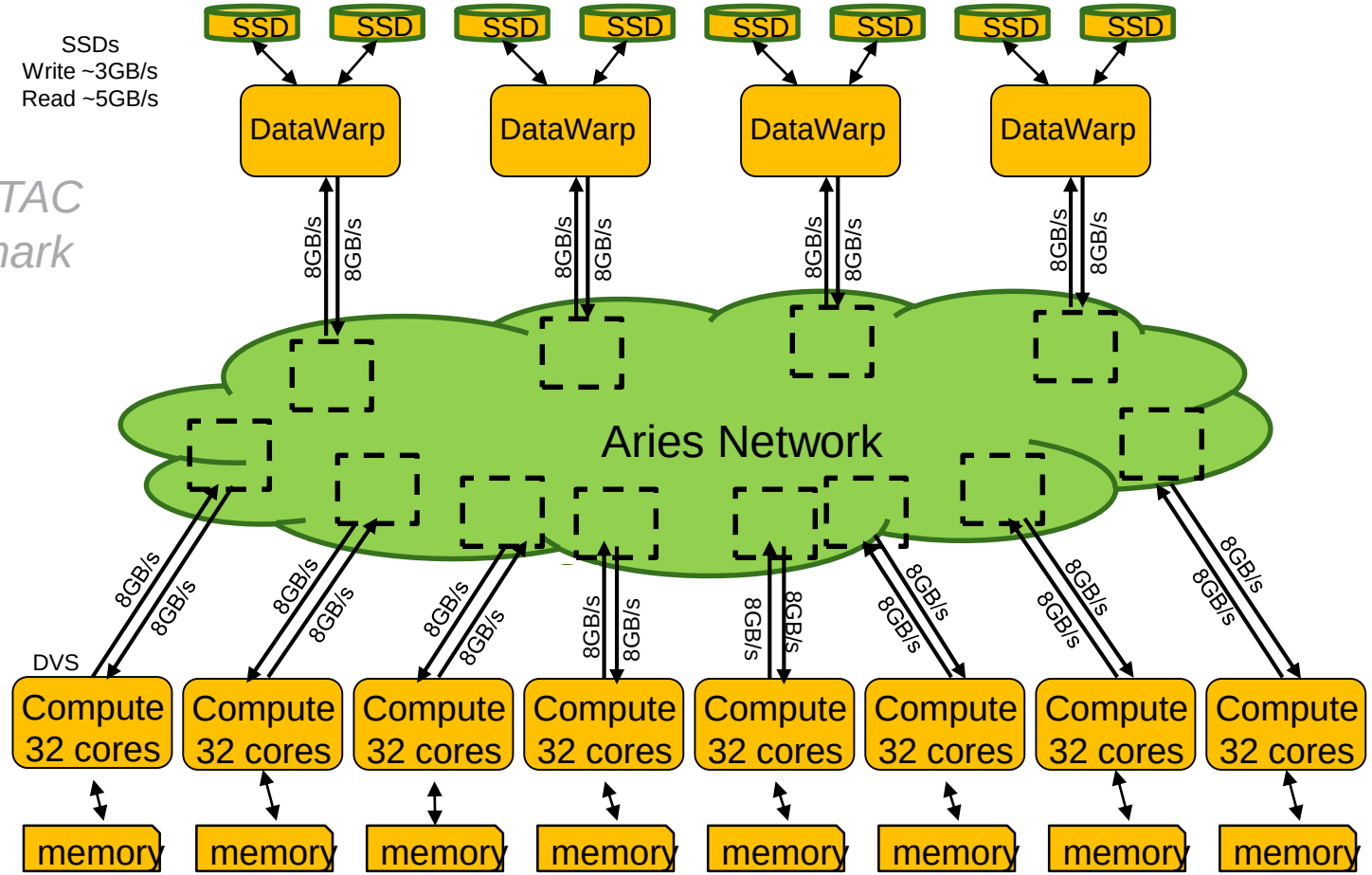
- **Definition**
  - A high-bandwidth, lower-capacity, “buffer” space
  - Backed by a disk-based PFS
- **Increased BB bandwidth decreases time programs spend on I/O**
  - BB can interact with PFS before, during, and after program use
  - Stage data in to BB before computes allocated
  - Stage data back out to PFS after computes deallocated
  - Stage data in or out while program in computational phase
- **Burst buffers offer improved bandwidth per dollar**
  - do faster I/O to BB, write out to slower PFS over time



DataWarp Notion – Minimize Compute Residence Time

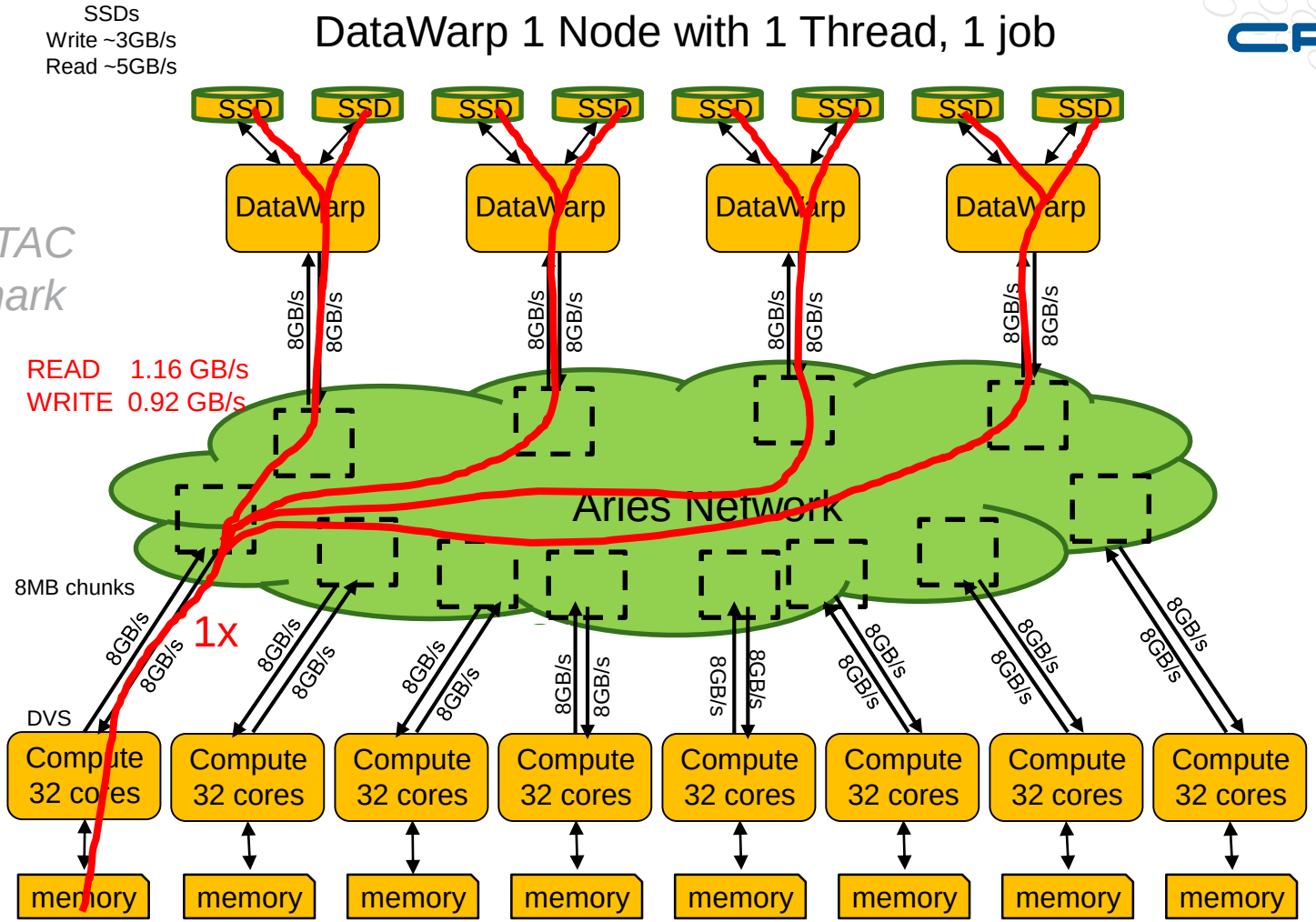


XC40 Architecture



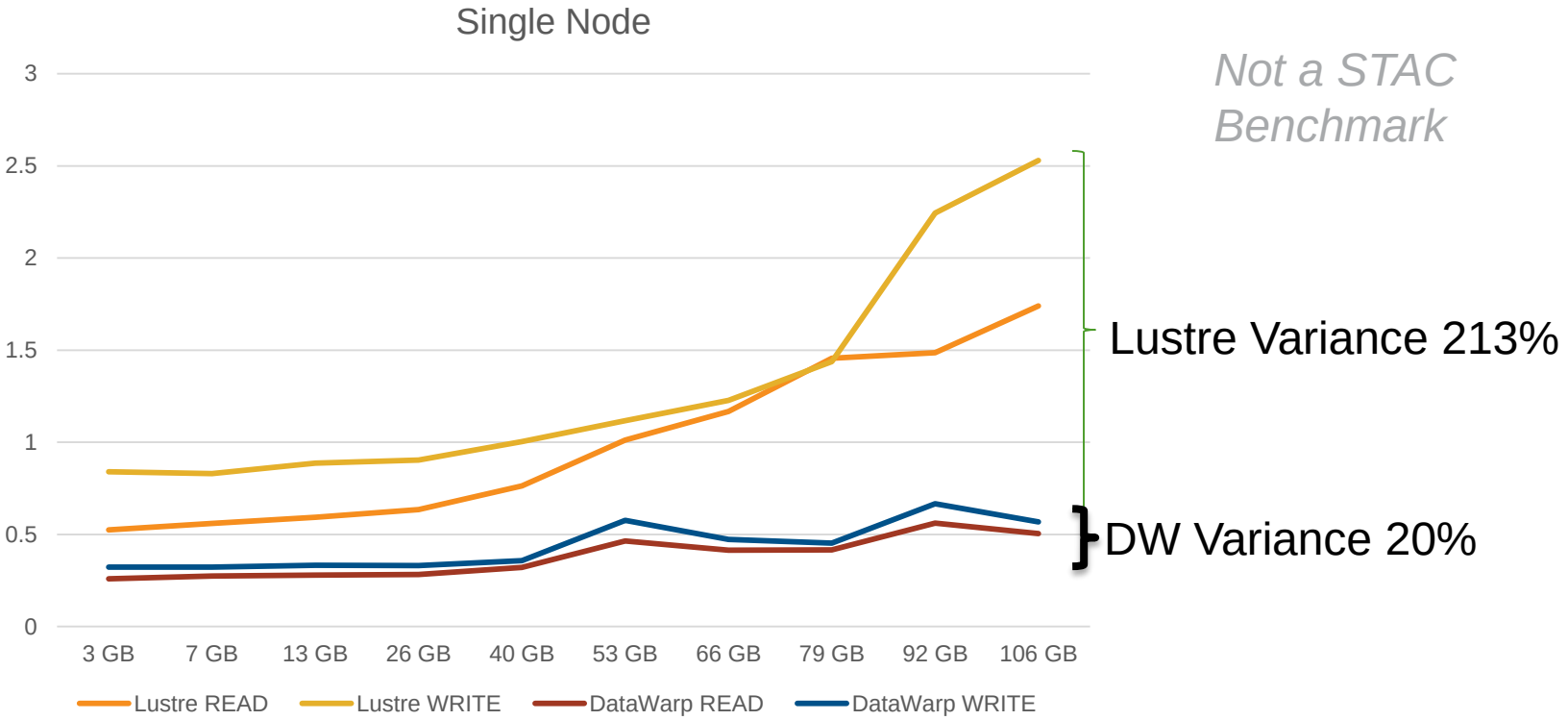
Not a STAC Benchmark

DataWarp 1 Node with 1 Thread, 1 job



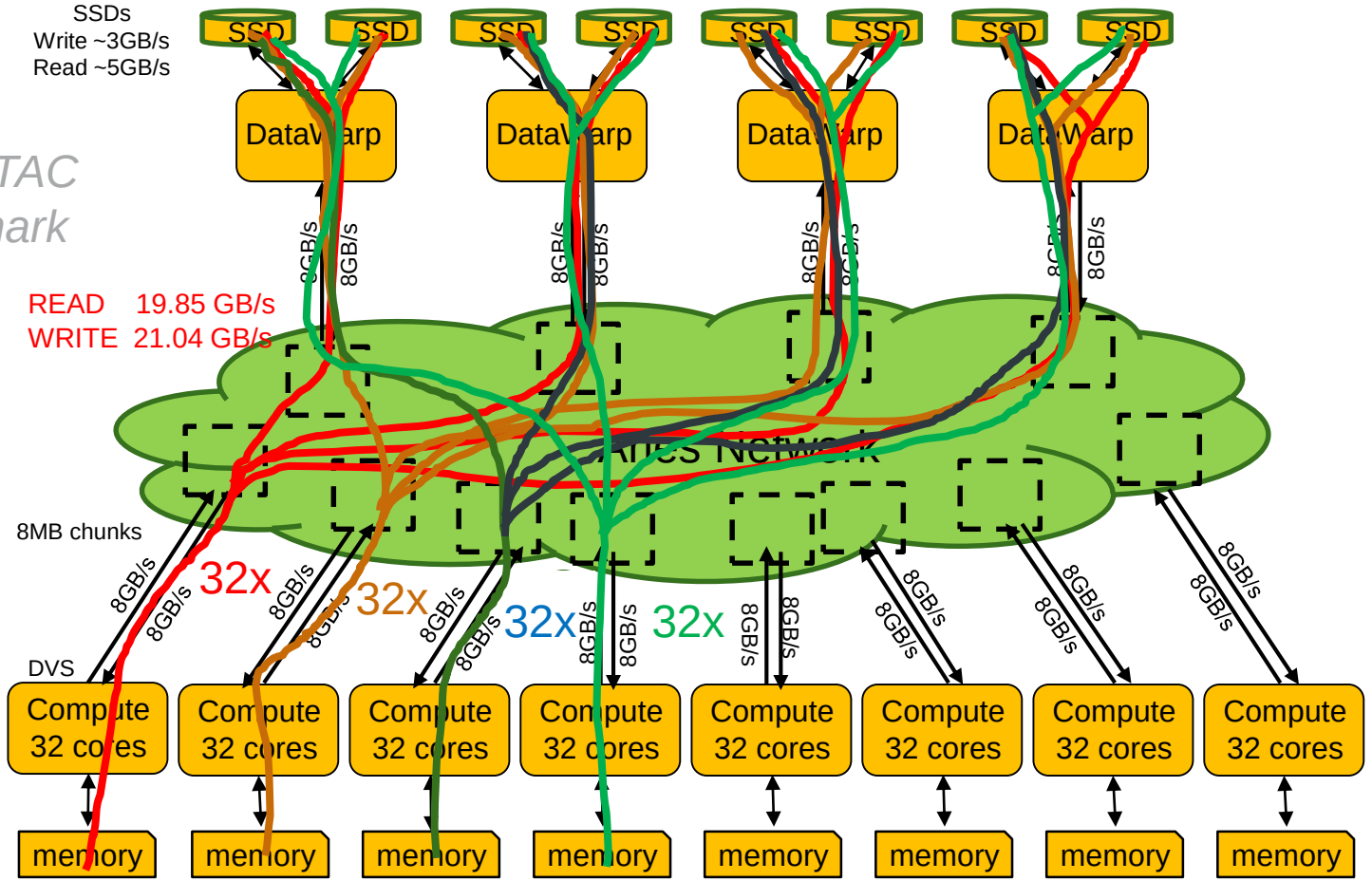
Metrics

DataWarp 51-78% faster than Lustre (R/W)



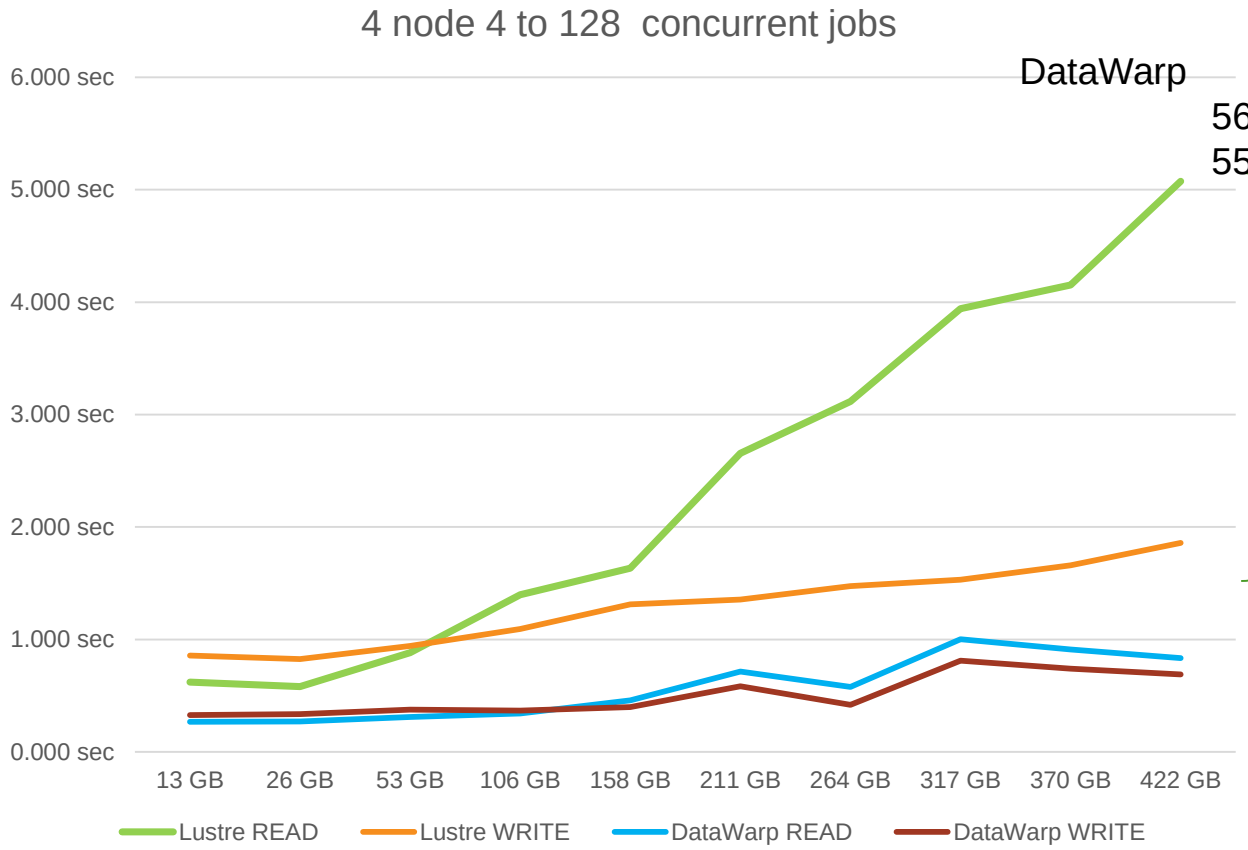
Not a STAC Benchmark

DataWarp 4 Nodes with 32 Threads, 128 jobs



Not a STAC Benchmark

Yet More Metrics



DataWarp

56→83.5% lower latency READS
55-71% lower latency WRITES

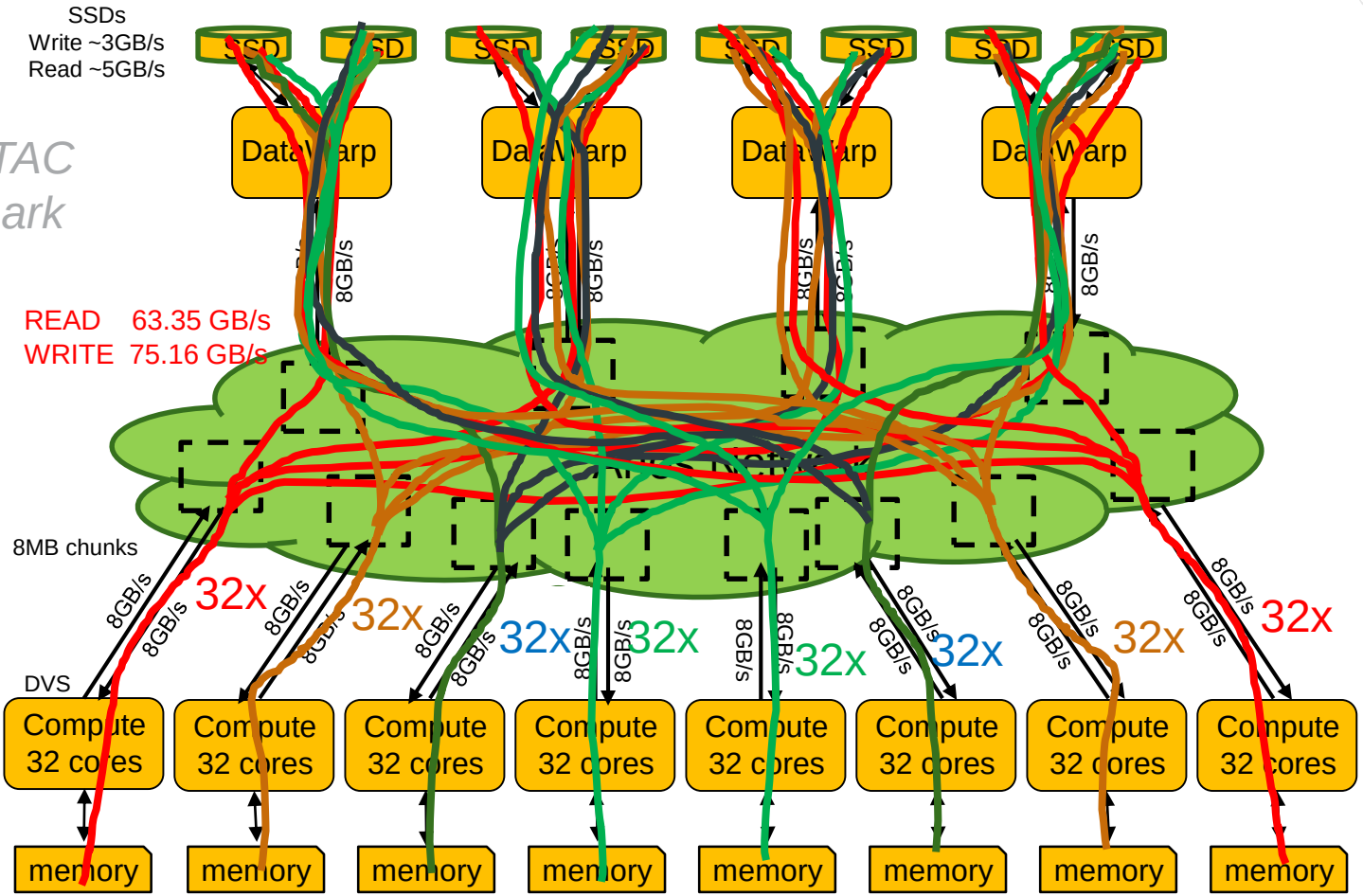
Lustre Variance 722%

Not a STAC Benchmark

DW Variance 111%

DataWarp 8 Nodes with 32 Threads, 256 jobs

Not a STAC Benchmark



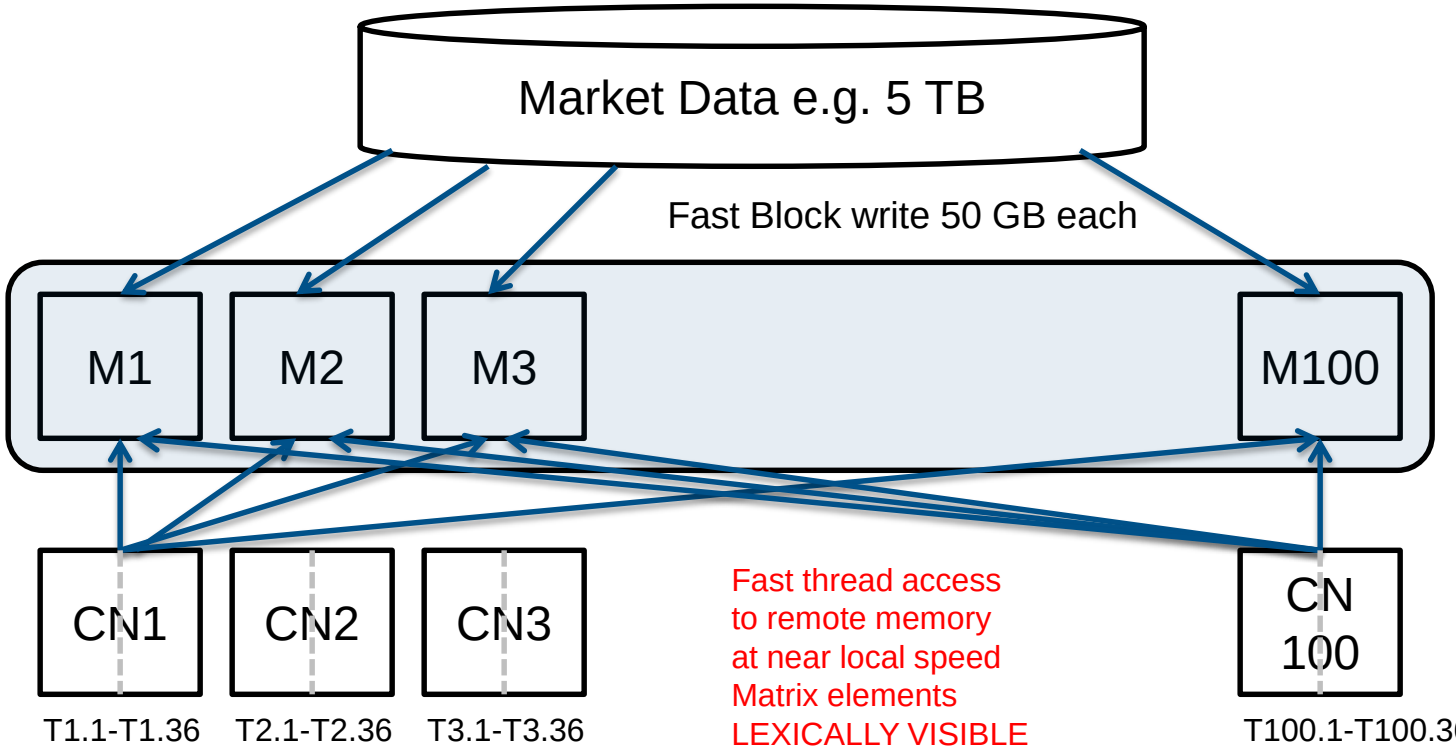
Great but not a Game-Changer

- Faster than PFS
- More effective than local SSD
- More manageable… BUT how to change the game….

Why not read shared data into memory?

- **There's lots of nodes – maybe 100Tb of memory across a cluster**
  - More than enough for all your market data
  - What if all nodes could access all the data at memory speeds….?
  - PERFORMANCE AND PRODUCTIVITY
- **Well you CAN**
  - Open SHMEM
  - PGAS languages e.g. Co-Array C++, UPC

Shared Global Memory

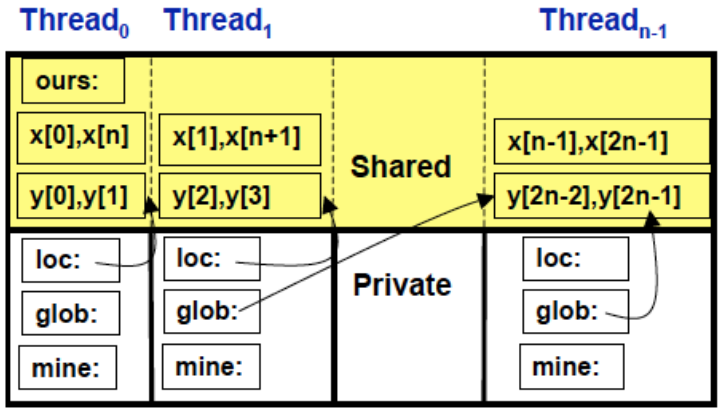


HPC Programming Model Taxonomy

- **Communication Libraries**
  - **MPI**, PVM, SHMEM, ARMCI, GASNet, …
- **Shared Memory Programming Models**
  - **OpenMP**, pthreads, …
- **Hybrid Models**
  - **MPI+OpenMP**, MPI+CUDA, MPI+OpenCL, …
- **Traditional PGAS Languages**
  - **Unified Parallel C (UPC)**, **Co-Array Fortran (CAF)**, Titanium (Java), Co-Array C++
- **HPCS Languages**
  - **Chapel**, **X10**, Fortress
- **GPU Programming Models**
  - **CUDA**, OpenCL, PGI annotations, CAPS, …
- **Others**
  - Global Arrays, Charm++, ParalleX, Cilk, TBB, PPL, parallel Matlabs, Star-P, PLINQ, Map-Reduce, DPJ, Yada, …

PGAS Programming Models

- **Characteristics:**
  - execute an SPMD program (Single Program, Multiple Data)
  - all binaries share a namespace
    - namespace is partitioned, permitting reasoning about locality
    - binaries also have a local, private namespace
  - compiler introduces communication to satisfy remote references
    - Cray compilers optimize by overlapping compute and communications, unlike e.g. OpenMP



Traditional PGAS Languages: in a Nutshell

Comparable performance to MPI, sometimes better

Co-Array C++: extend C++ by adding…

- a new array dimension to refer to processor space
- collectives and synchronization routines

UPC: extend C by adding support for…

- block-cyclic distributed arrays
- pointers to variables on remote nodes
- a memory consistency model

Titanium: extend Java by adding support for…

- multidimensional arrays
- pointers to variables on remote nodes
- synchronization safety via the type system
- …region-based memory management
- …features to help with halo communications and other array idioms