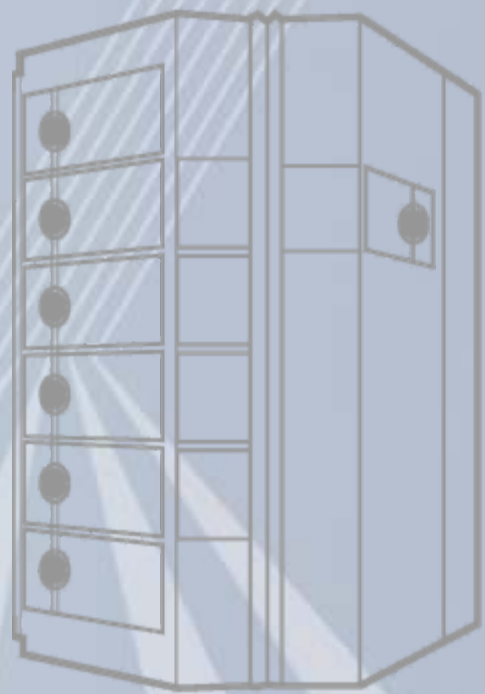
BlackWidow System Update

CRAY
THE SUPERCOMPUTER COMPANY

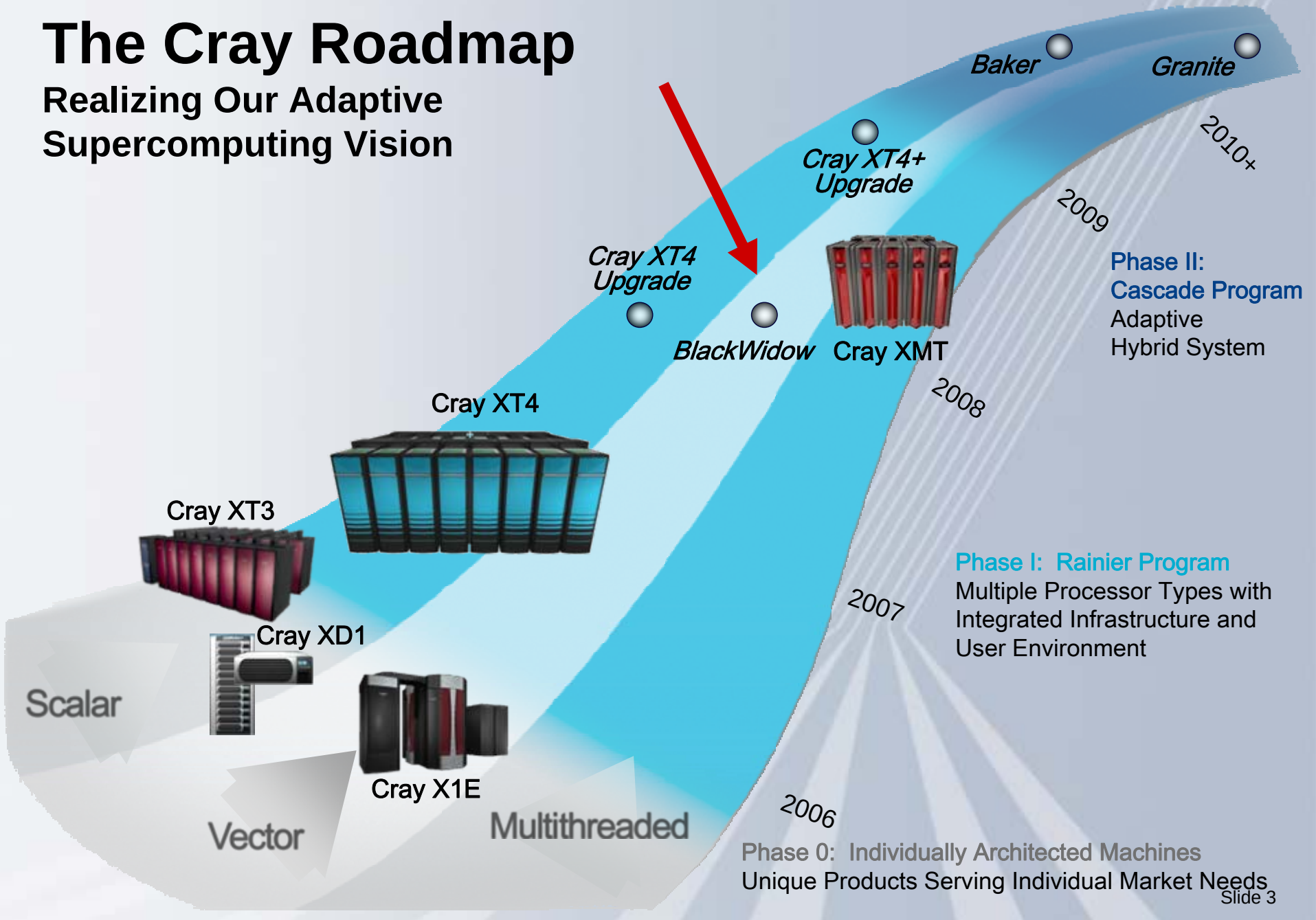
BlackWidow

- System Update
- Performance Update
- Programming Environment
- Scalar & Vector Computing
- Roadmap



The Cray Roadmap

Realizing Our Adaptive Supercomputing Vision



BlackWidow Summary

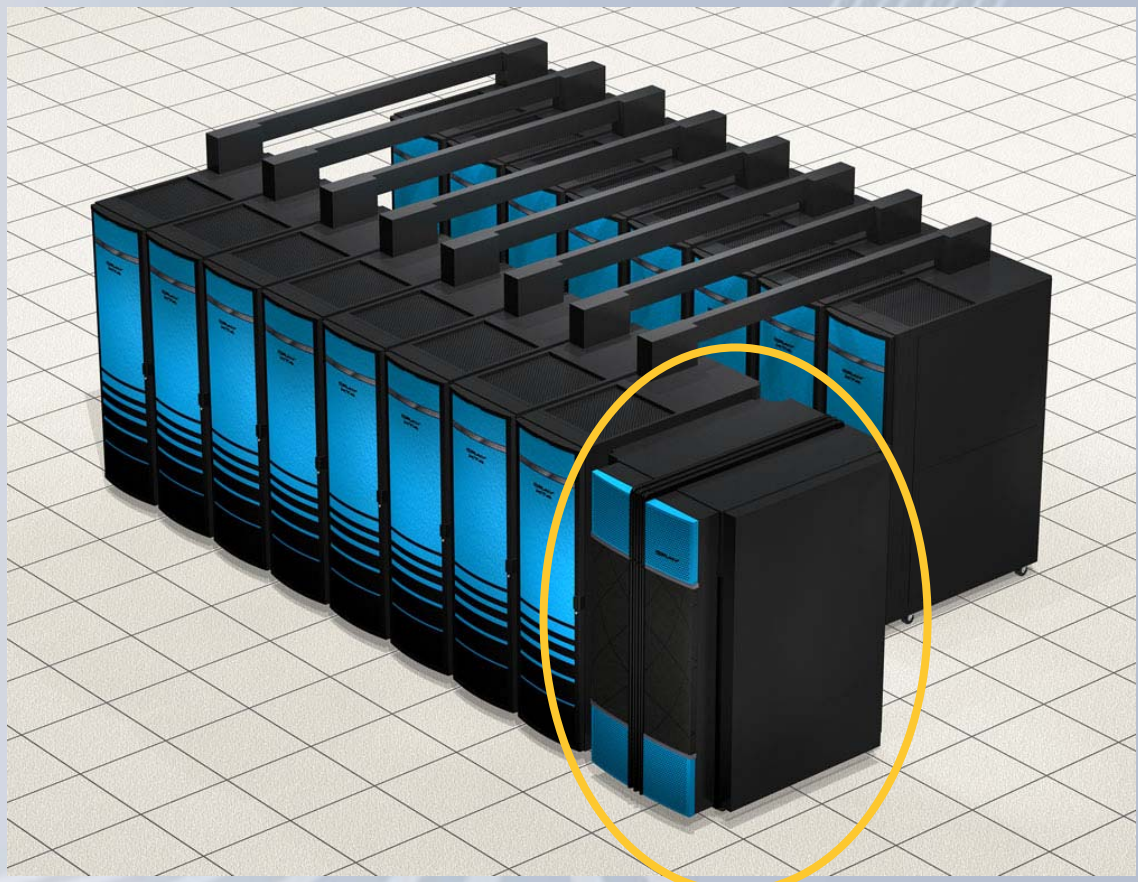
- Project name for Cray's next generation vector system
- Follow-on to Cray X1 and Cray X1E vector systems
- Special focus on improving scalar performance
- Significantly improved price-performance over earlier systems
- Closely integrated with Cray XT infrastructure
- Product will be launched and shipments will begin towards the end of this year

"Vector MPP" System

- **Highly scalable – to thousands of CPUs**
- **High network bandwidth**
- **Minimal OS "jitter" on compute blades**

BlackWidow System Overview

- Tightly integrated with Cray XT3 or Cray XT4 system
- Cray XT SIO nodes provide I/O and login support for BlackWidow
- Users can easily run on either or both vector and scalar compute blades



BlackWidow cabinets attached to Cray XT4 system (artist's conception)

Compute Cabinet

Blower

Rank1 router modules
- Up to 4 per chassis

Chassis
- 2 per cabinet

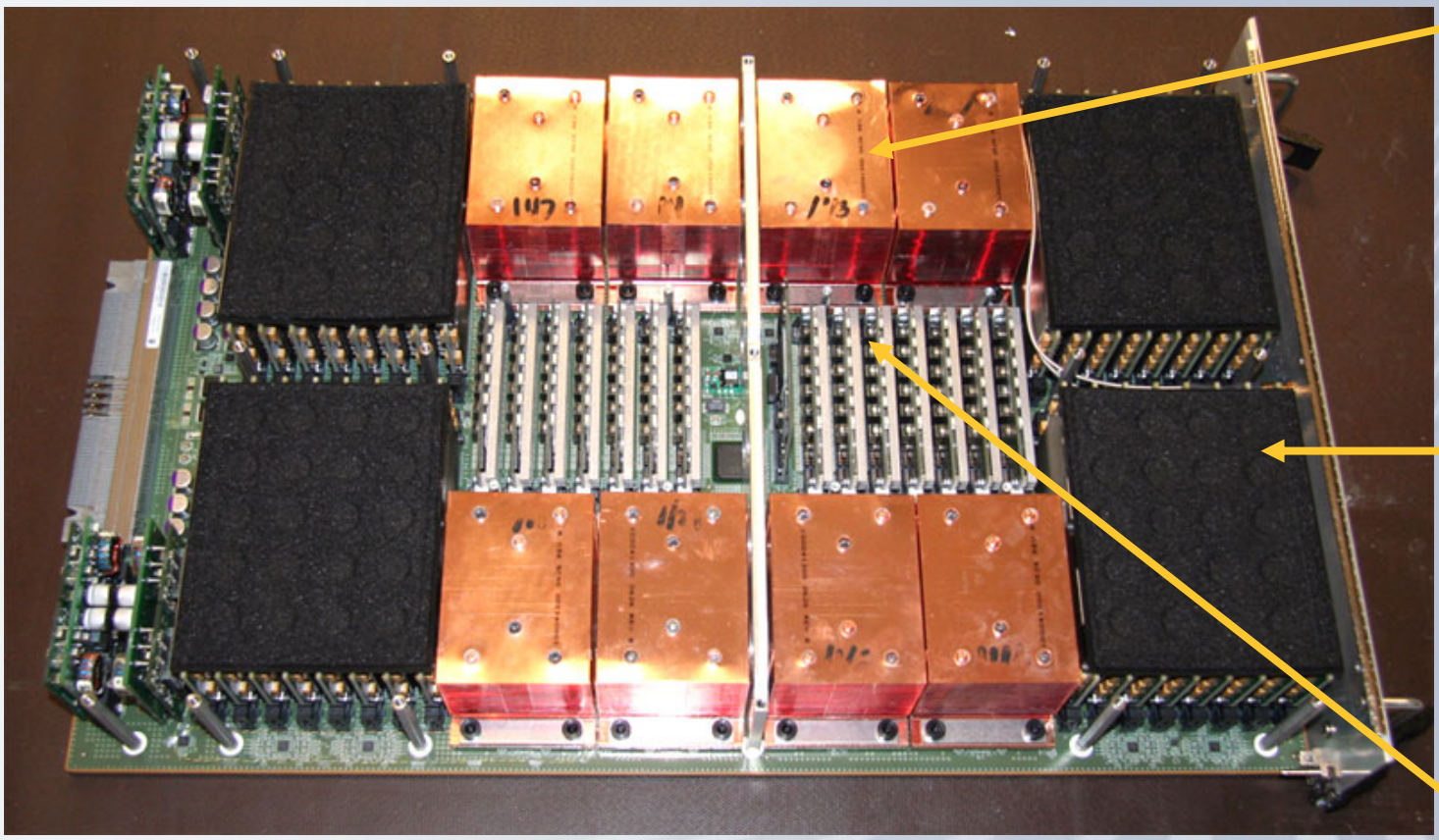
Compute and bridge blades
- Up to 8 per chassis

Power supplies

Back

Front

Cray BlackWidow Compute Blade



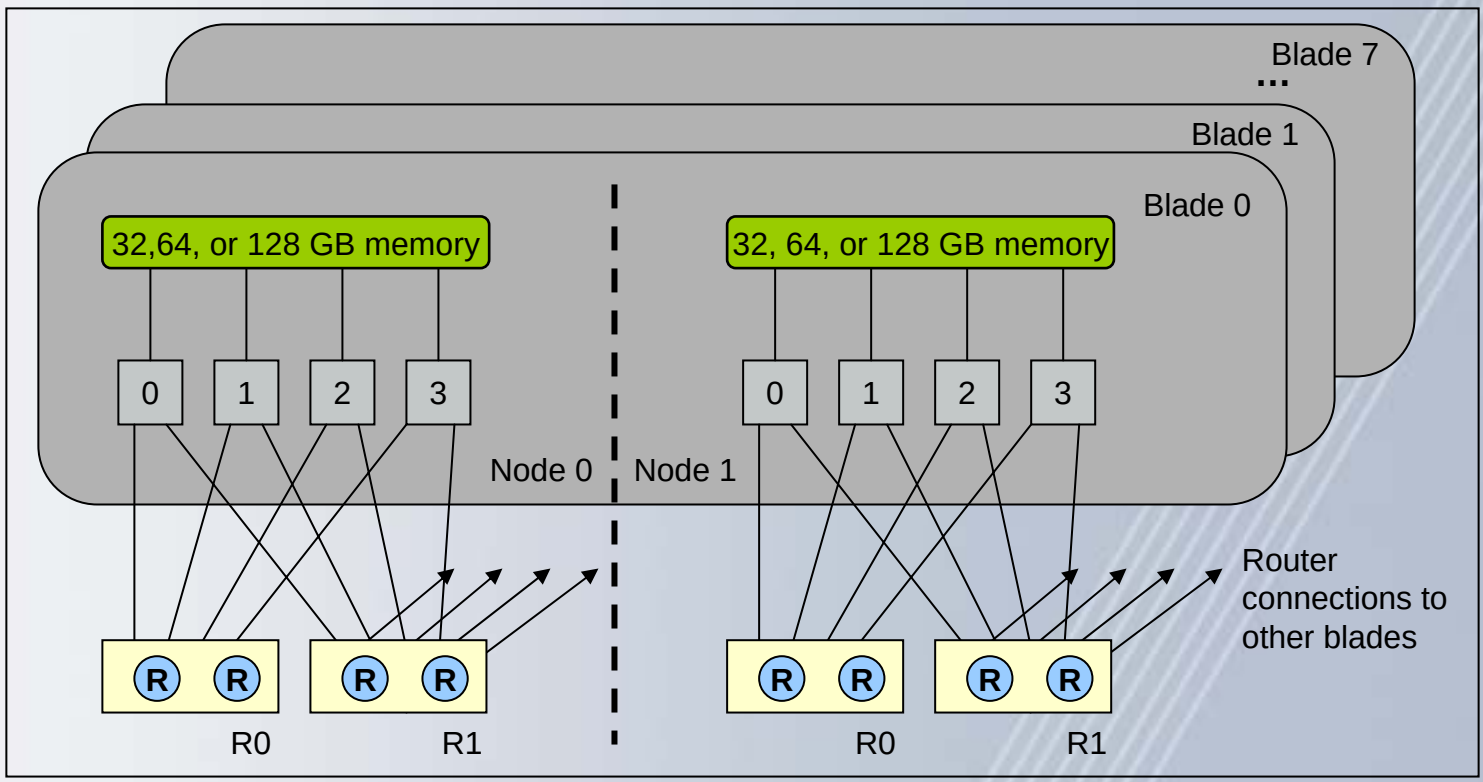
BlackWidow vector CPUs
- 8 per blade
- Configured as two 4-way SMP nodes

Memory
- 32, 64, or 128 GB per node

Voltage regulator modules

22.8” x 14.4”

BlackWidow Chassis



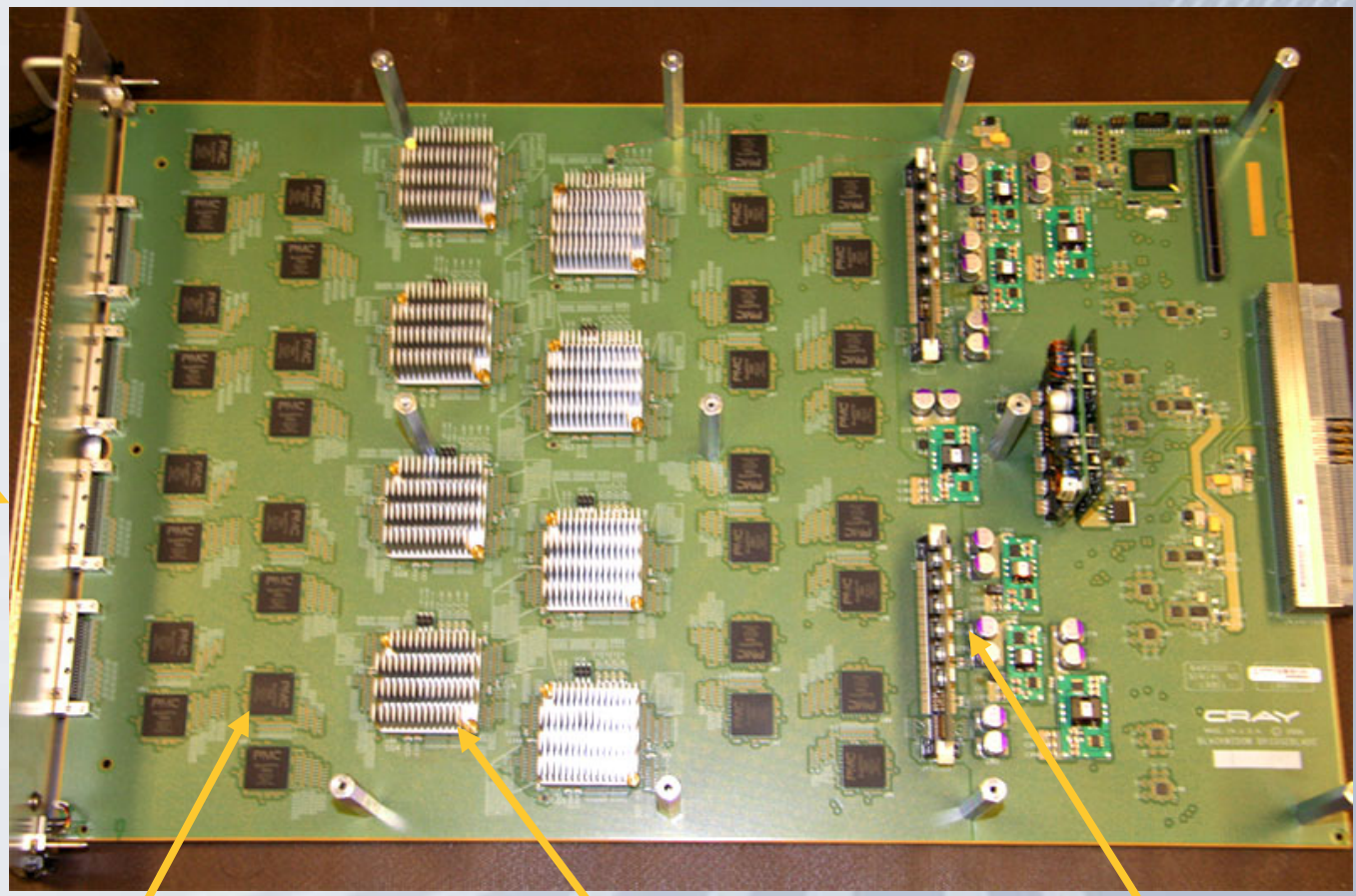
Eight blades per chassis

- 64 CPUs configured as 16 4-way SMP nodes

Two chassis per cabinet

- 128 CPUs per cabinet

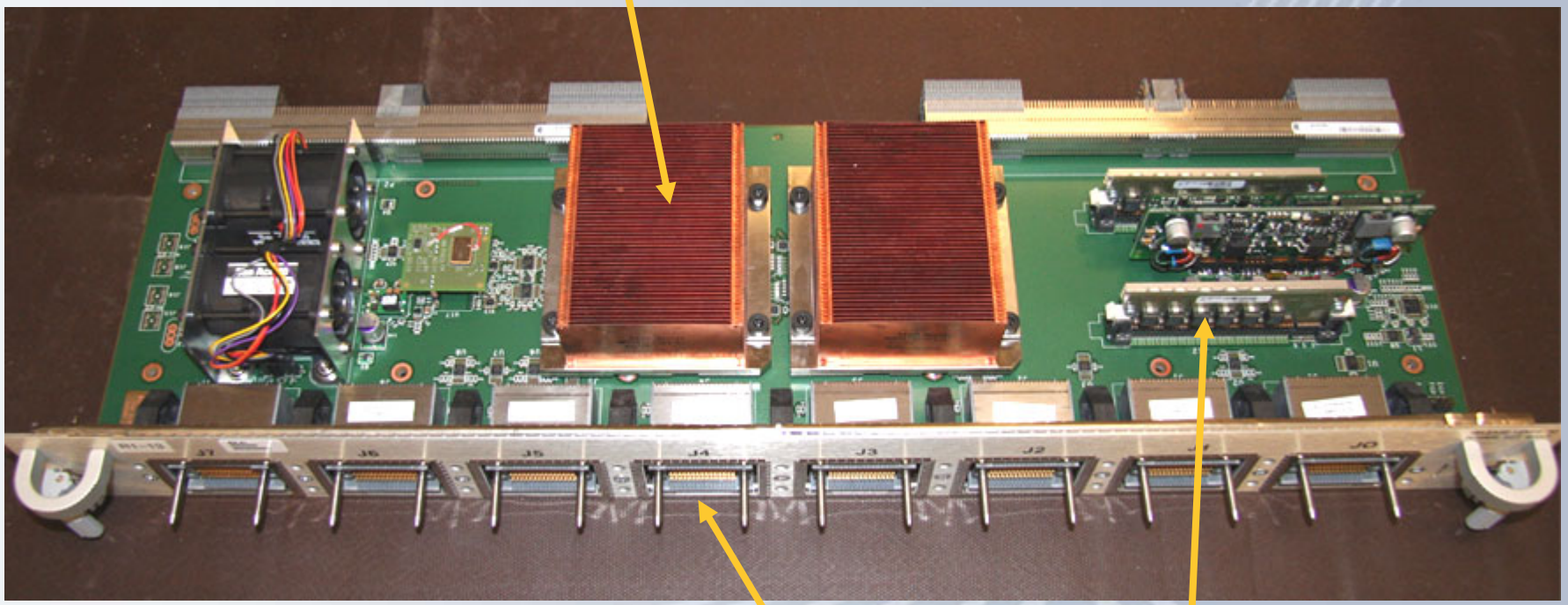
Bridge Blade



Rank1 Router Module

YARC chips
•2 per module

23” x 7”

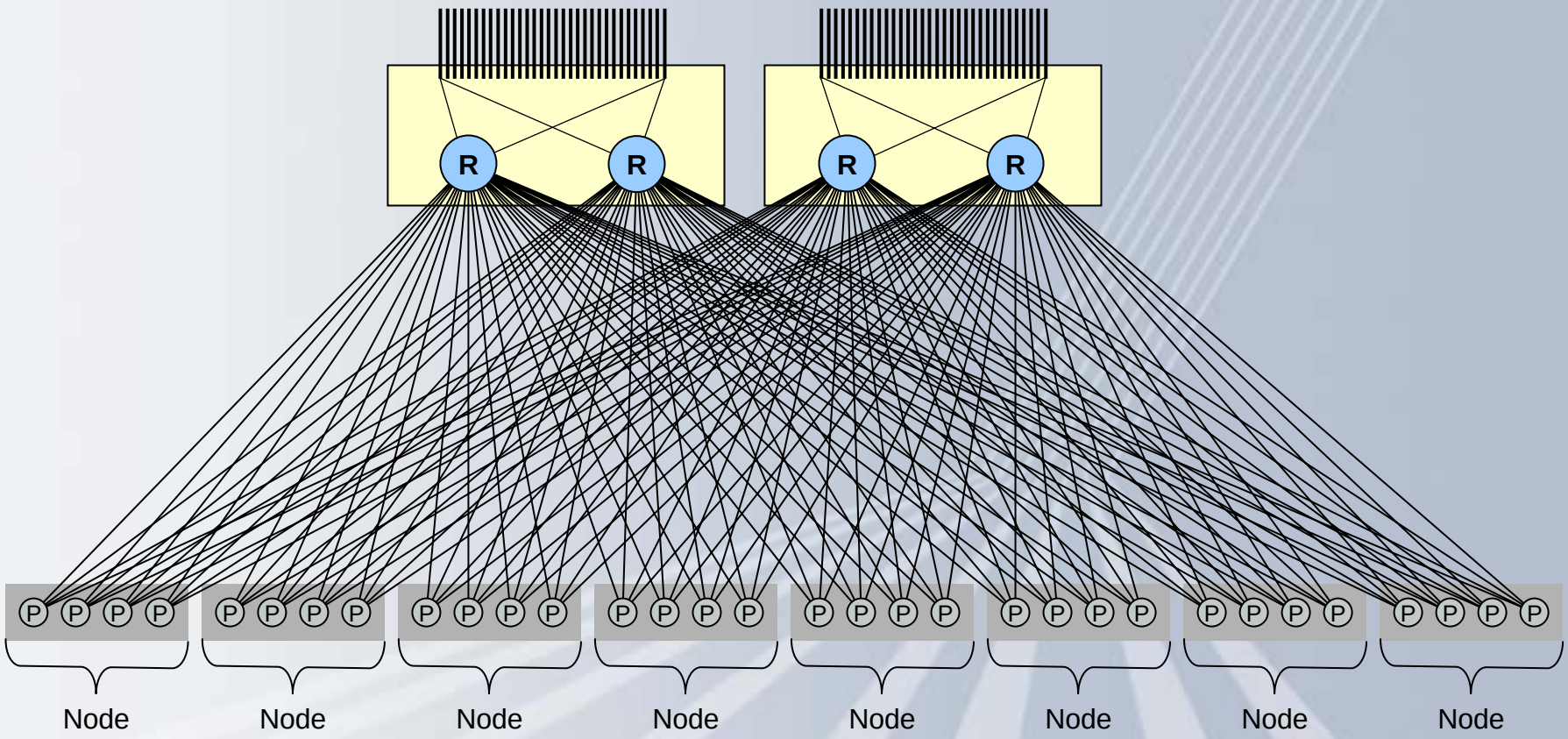


Network cable connectors

Voltage regulator modules

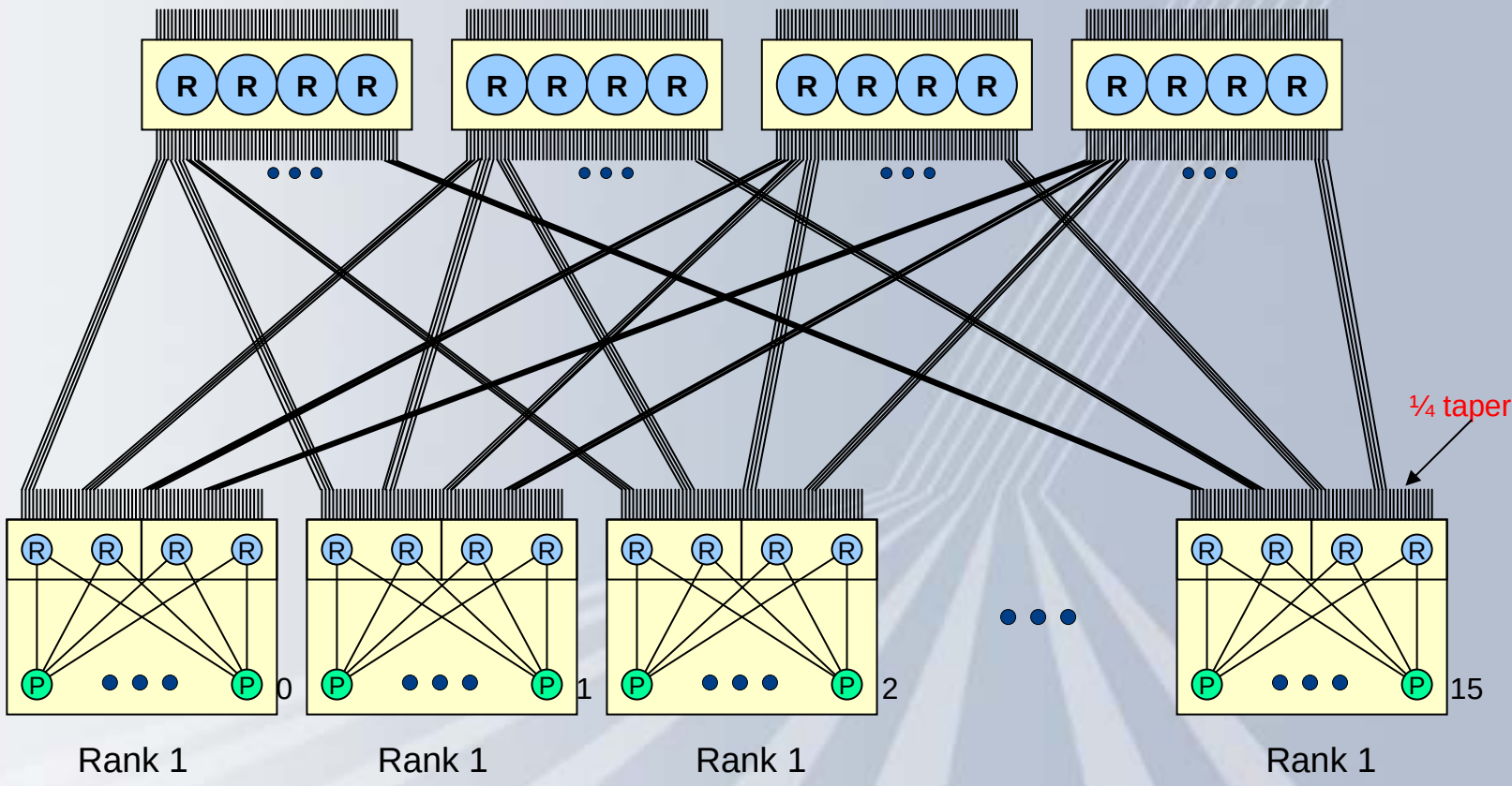
Fat Tree Network Scales to 1000s of CPUs

- Rank 1 building block – 32 CPUs



Fat Tree Network Scales to 1000s of CPUs

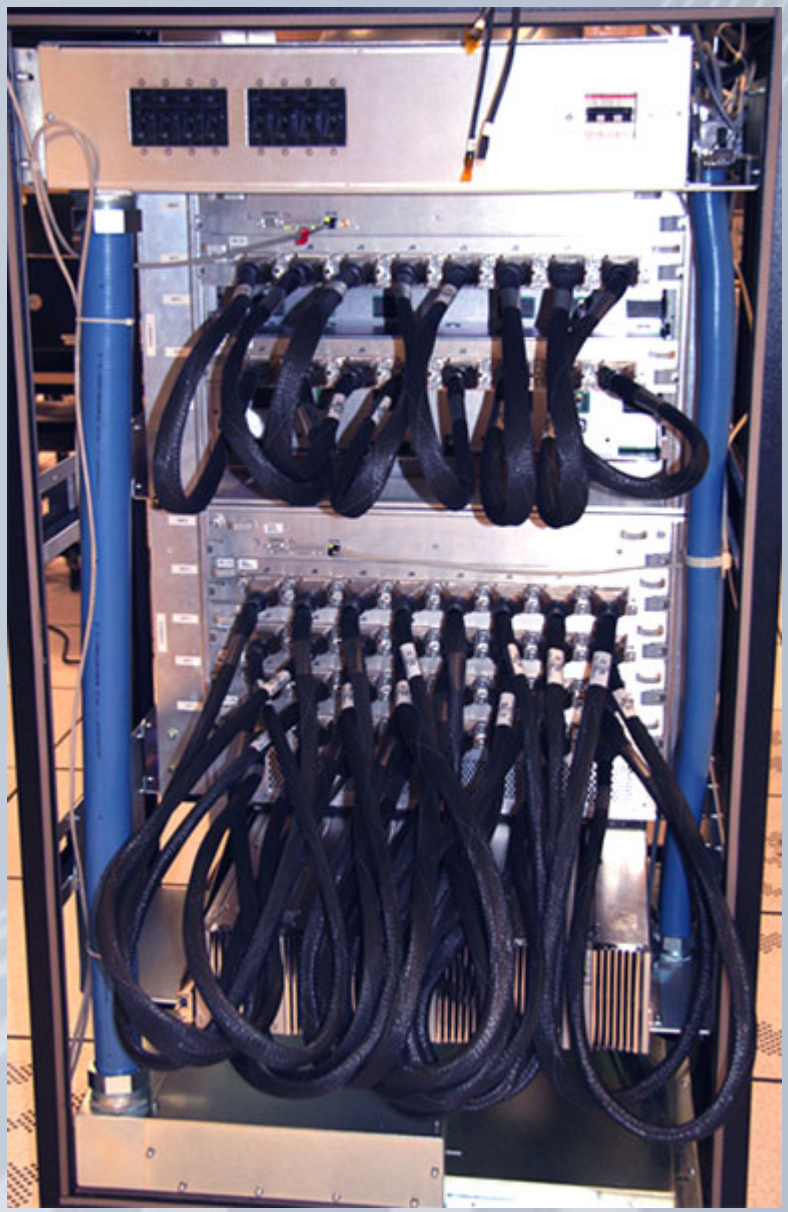
- Rank 2 building block – up to 1024 processors
  - Diagram shows 512 processors (4 compute cabinets)



Compute Cabinet

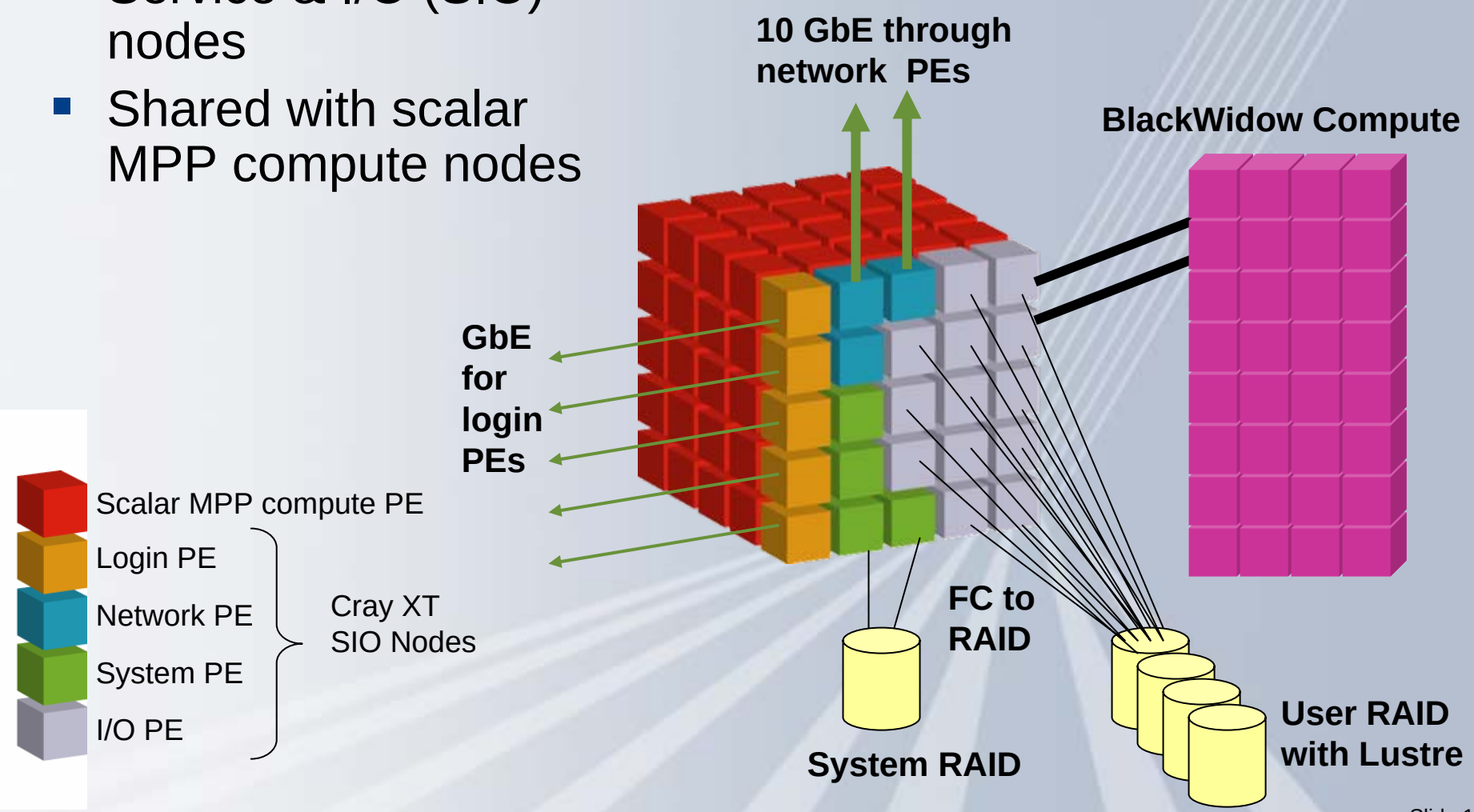
Back of cabinet showing cables connecting Rank1 router modules

- 2 modules in top chassis
- 4 modules in bottom chassis



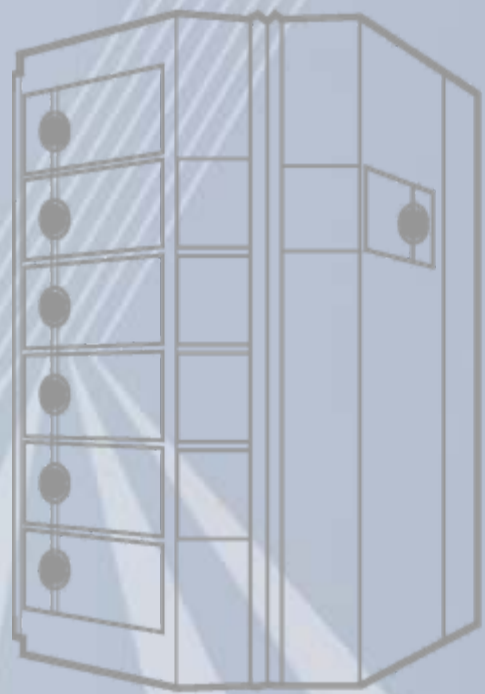
I/O and Networking

- Delivered by Cray XT Service & I/O (SIO) nodes
- Shared with scalar MPP compute nodes



BlackWidow

- System Update
- Performance Update
- Programming Environment
- Scalar & Vector Computing
- Roadmap



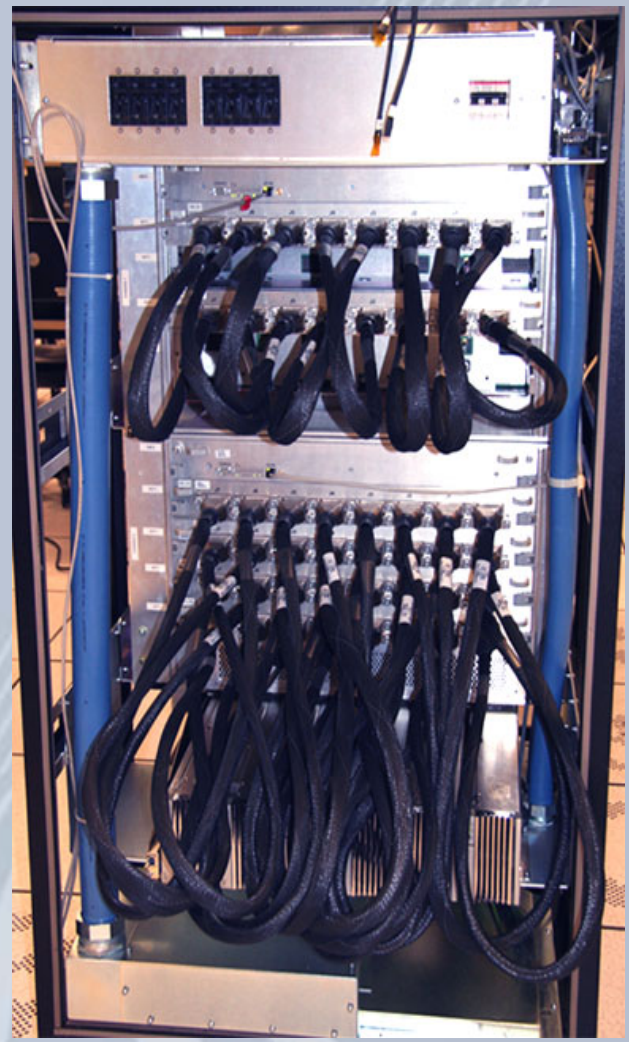
Key Comparisons with Cray X1E

	Cray X1E	BlackWidow
Processor	Dual-core, MSP	Single-core, single-CPU (no MSP!)
Processor performance	4.5 GFLOPS (SSP)	20+ GFLOPS
Memory bandwidth	34 GB/s/MSP	41 GB/s/CPU
Network bandwidth	6.4 GB/s/MSP	8 GB/s/CPU
Price/MFLOP	~$5/MFLOP	$1/MFLOP
Cabinets	Two – air-cooled, liquid-cooled	One – air cooled (with liquid-cooled upgrade option)
I/O Infrastructure	Unique Cray-developed components	Cray XT SIO nodes and software
Programming environment	Same as BlackWidow	Same as Cray X1E
Operating system	Irix-based	Linux-based (CNL)
User environment	Cray X1E CPU, CPES	Cray XT login nodes
File system	XFS	Lustre

Early Benchmark Results

Results are on prototype BlackWidow hardware

- Some performance limitations; e.g. speculative loads are disabled on prototype hardware
- Comparisons are to X1 (i.e. not X1E) performance



Key Metrics

- HPL (Linpack)
  - Currently estimate BlackWidow will get 18+ GFLOPS/CPU
    - 90+% of peak

- STREAM Triad Estimates
  - Triad

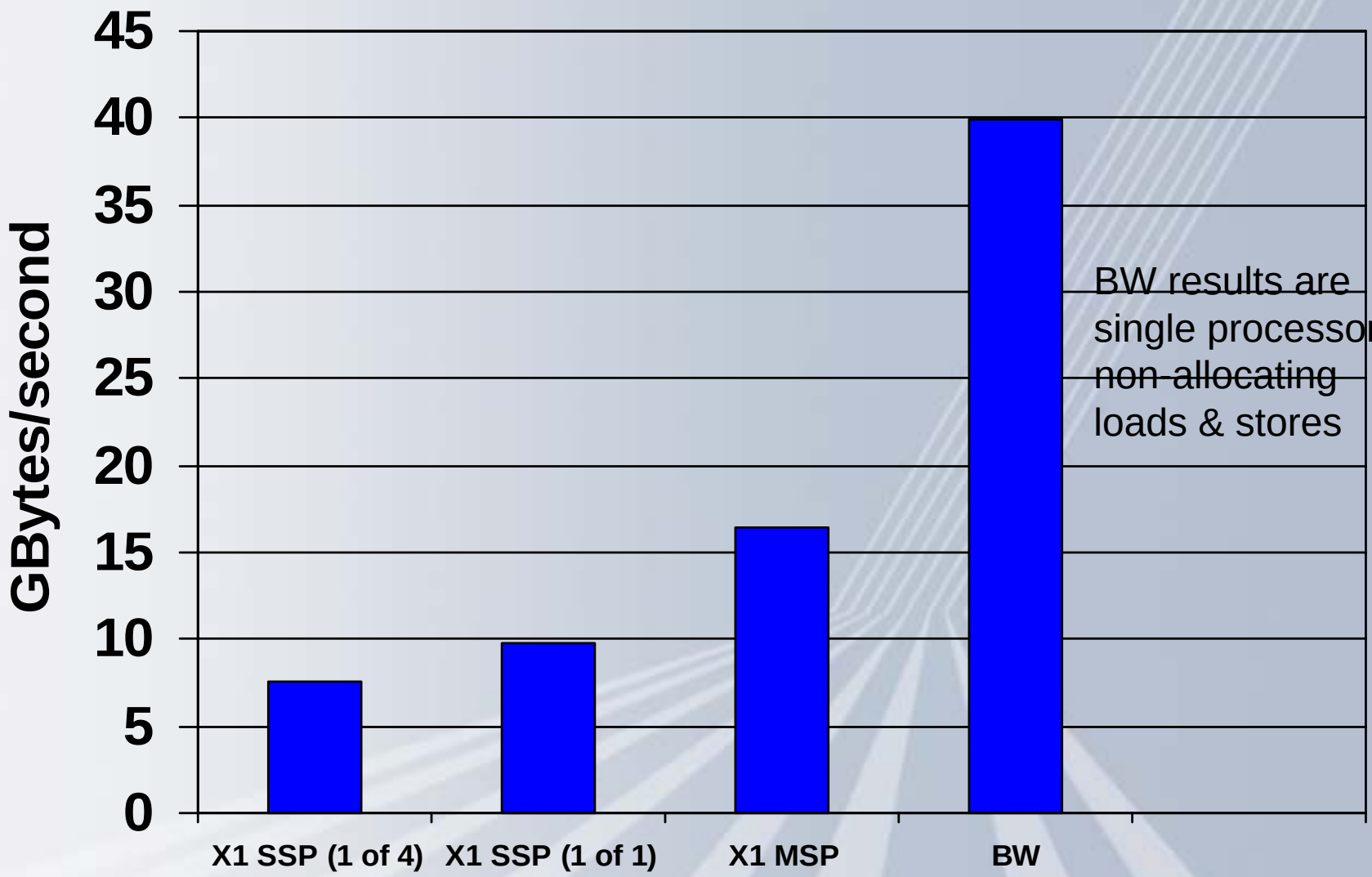
CPUs	GB/s
1	39
2	83
4	121

Triad (stride = 4 – no cache locality)

CPUs	GB/s
1	27
2	44
4	40

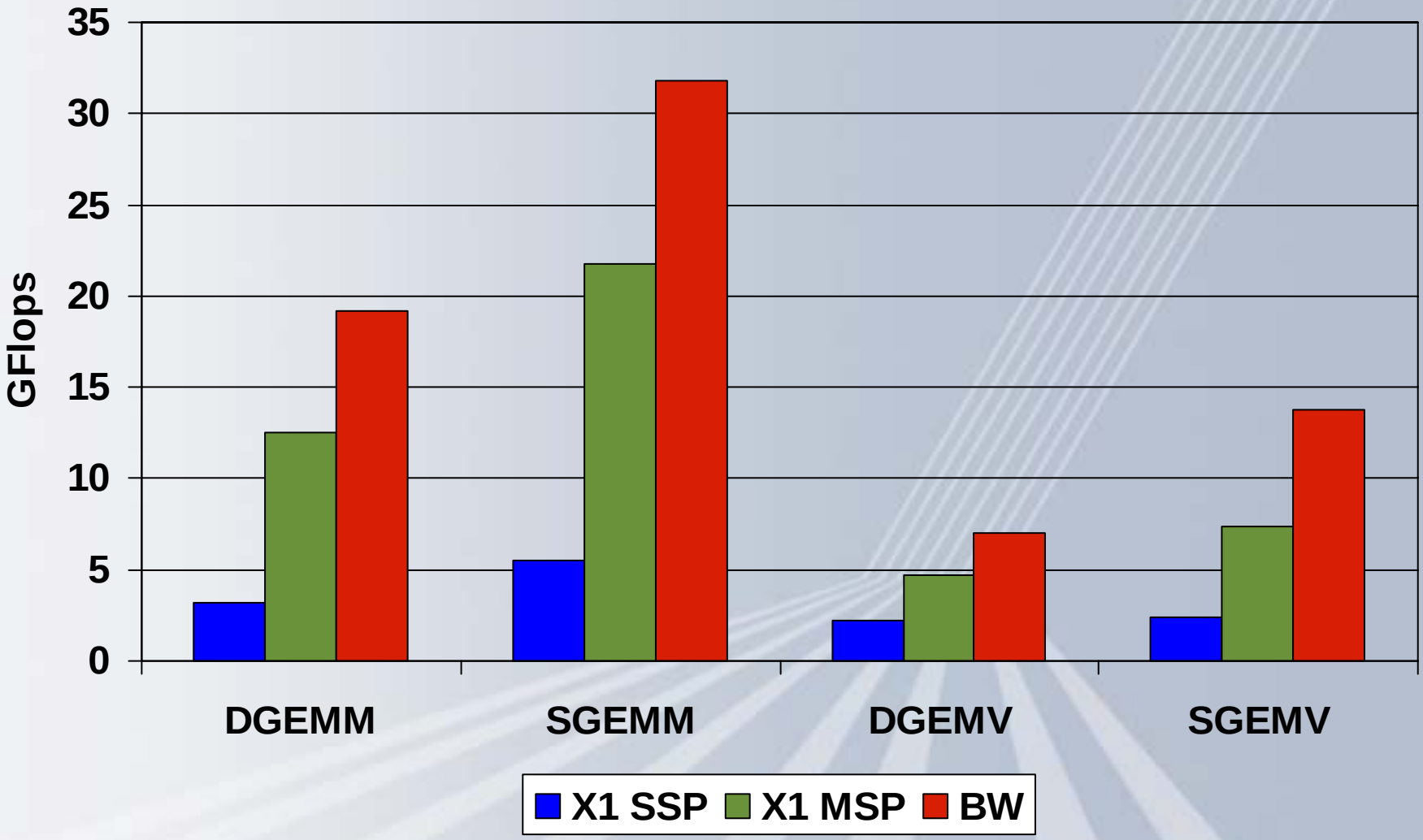
- MPI Metric
  - Estimated latency for zero-length message: 4.9 microseconds

STREAM Triad Performance Comparison

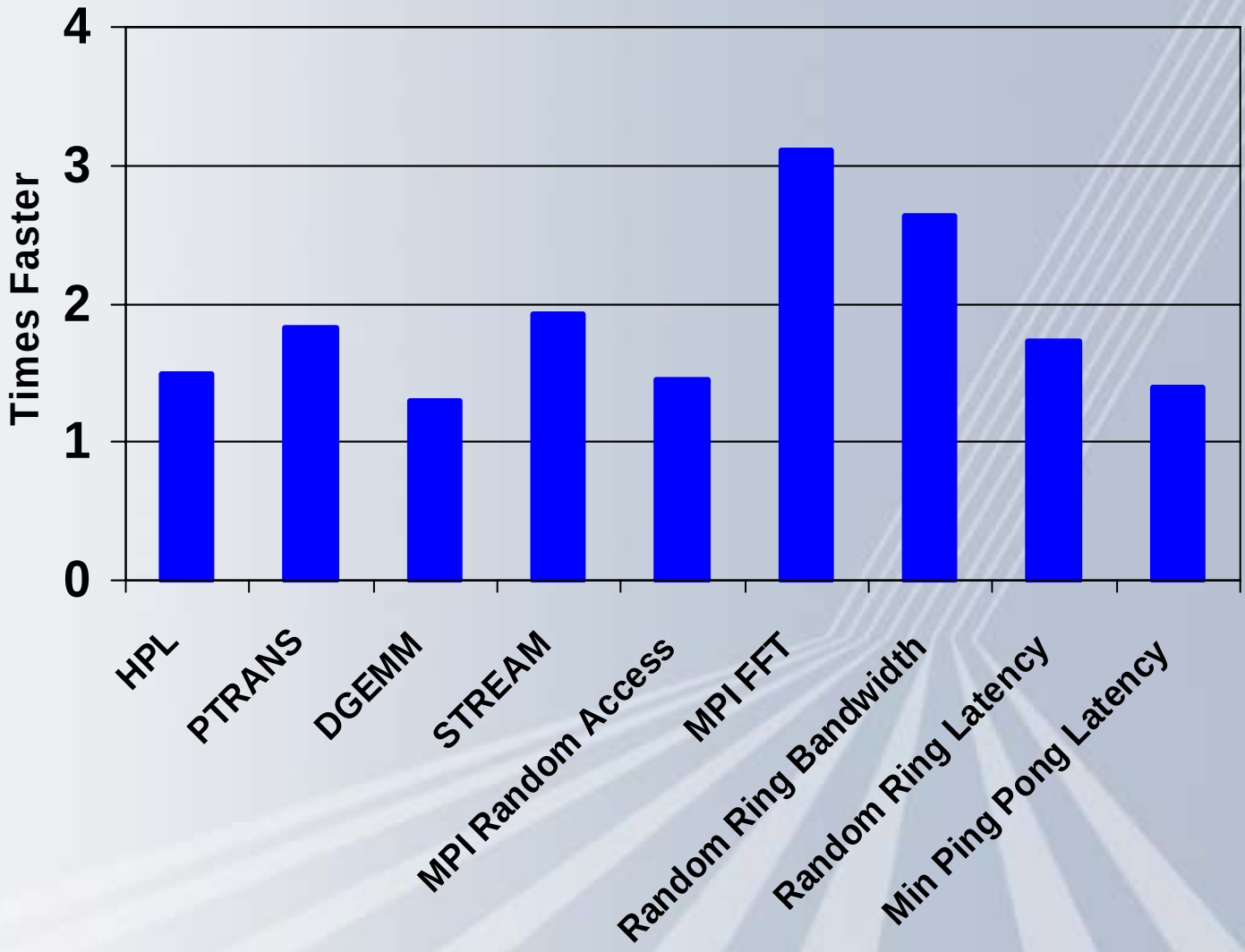


DGEMM, SGEMM, DGEMV, SGEMV

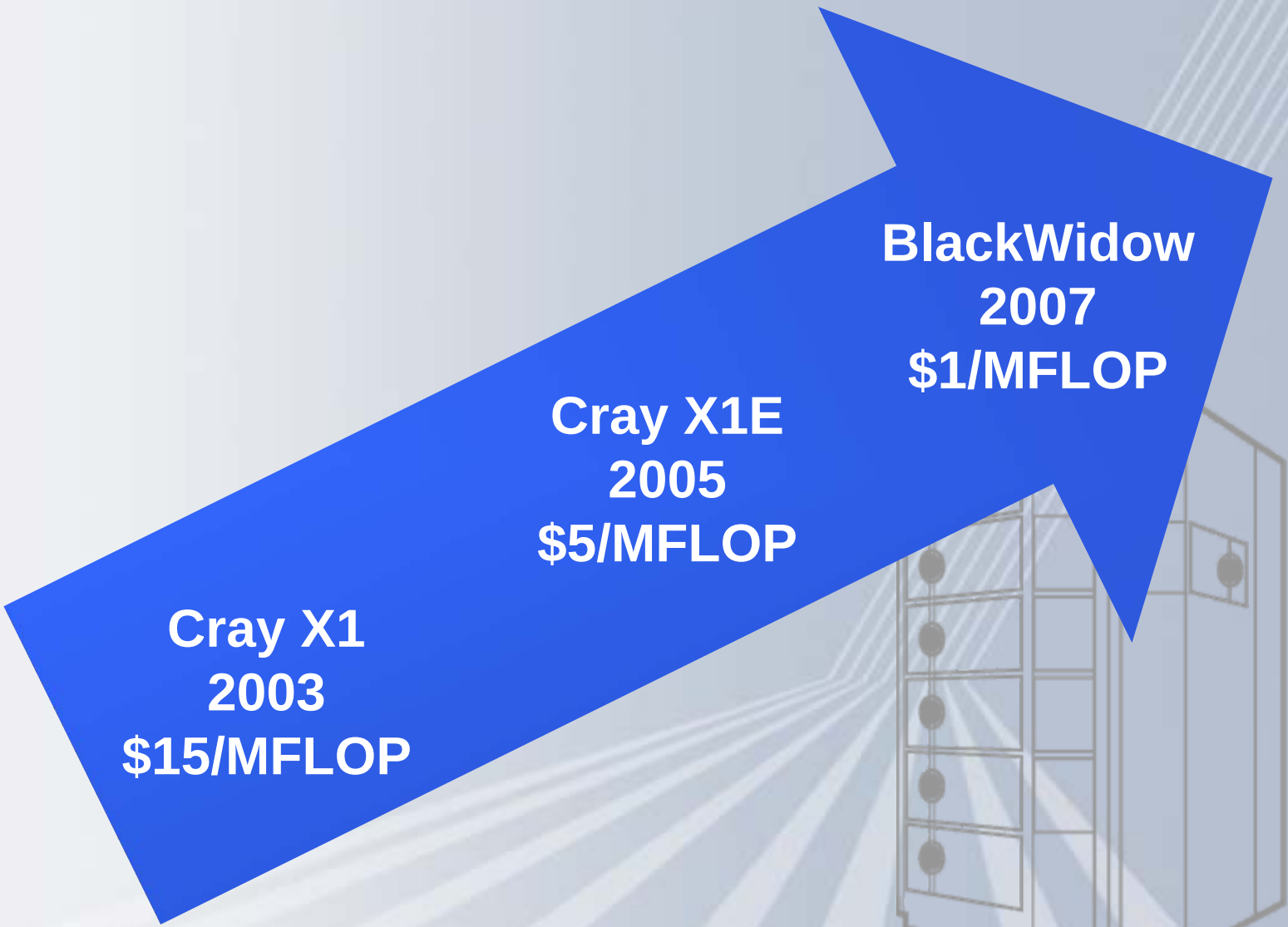
(Compiler generated user level Fortran routines, N=5000).



HPCC Performance (4 BW PEs vs 4 X1 MSPs)

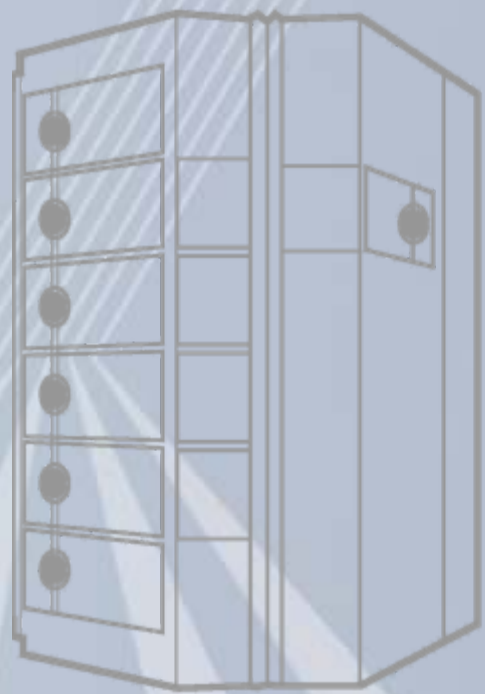


BlackWidow Price/Performance Gains



BlackWidow

- System Update
- Performance Update
- Programming Environment
- Scalar & Vector Computing
- Roadmap



Programming Environment

- BlackWidow has a single Programming Environment product
- Includes both Fortran and C/C++
- Also includes MPI and SHMEM libraries
  - No MPT product
- Licensing: base fee, fee per socket

- Compiles are performed on Cray XT login nodes

BlackWidow OS Features

- UNICOS/lc operating system is distributed between BlackWidow compute nodes and Cray XT SIO nodes
- BlackWidow compute nodes:
  - Linux (SLES 10) kernel with minimum commands and packages
  - Lustre client
  - TCP/IP Networking support
  - Dump/crash support
  - HSS/RAS support and resiliency
  - Application monitoring and management
  - Linux process accounting, application time reporting
- Cray XT SIO nodes:
  - Linux (SLES 10) kernel and “full set” of commands and packages on login nodes
  - Other SIO nodes run Lustre OSS and MDS, networking connections, scheduling and workload management software

Parallel Programming

- MPI-2
- Co-Array Fortran
- UPC 1.2
- OpenMP (within a node)
- Shmem
- Pthreads (within a node)
- MPMD (multiple binary launch)

Resource Scheduling

- ALPS resource scheduler
- Design goals
  - Manage scheduling and placement for distributed memory applications
  - Support predictable application performance
  - Conceal architecture specific details
  - Guarantee resource availability for batch and interactive requests
  - Integrate with and rely upon workload management applications
    - PBS Pro, LSF, etc.
  - Extensible, scalable, and maintainable
- Distributed design:
  - apinit (one per process) runs on BlackWidow compute nodes
  - All other daemons and commands run on Cray XT service nodes

TotalView Licensing Model

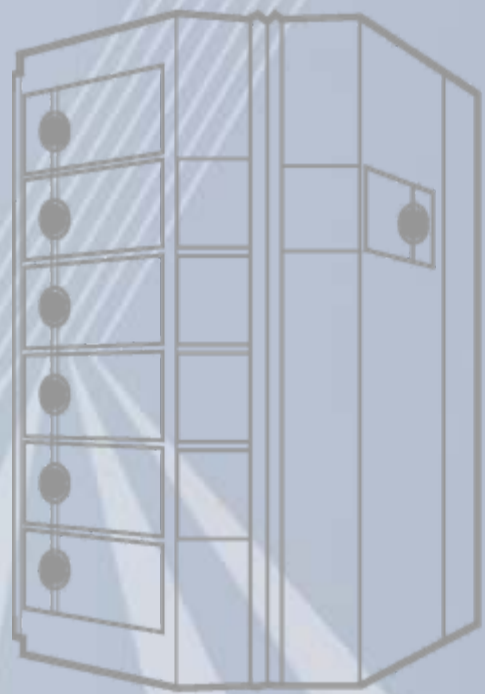
- Moving to a new licensing model for TotalView
- TotalView Technologies calls this “team licensing”
  - They still offer old model – “enterprise licensing”
- Customer buys “tokens”, where a token allows one MPI process to be debugged on one CPU
- Tokens are managed as a pool
- So 32 tokens can be used in a variety of ways:
  - One user debugging 32-process application
  - 32 users each debugging 1-process application
  - And everything in between!
- Easy to upgrade – just buy more tokens!
- See TotalView web site for details, more examples

BlackWidow Software Roadmap

- **New BlackWidow-specific capabilities including:**
  - Scheduling enhancements and improvements
  - New versions of Linux kernel for BlackWidow compute node
  - New versions of Lustre client software
  - New versions of PBS Pro and Totalview
- **BlackWidow customers will also benefit from new releases of XT SIO node software:**
  - New storage, networking, and device support
  - New and updated login node (Linux distribution) features
  - New versions of Lustre server software

BlackWidow

- System Update
- Performance Update
- Programming Environment
- Scalar & Vector Computing
- Roadmap



Scalar and Vector Processing at Boeing

Cray X1

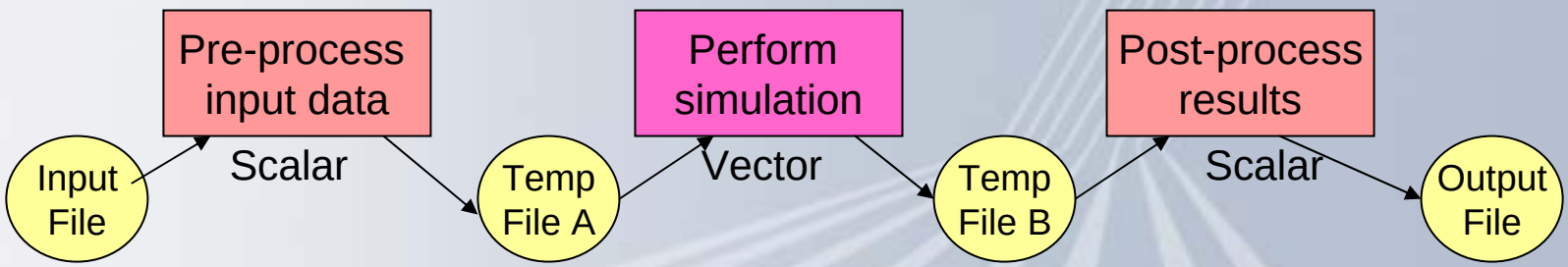
Overflow Wing Design Code

Linux Networx Clusters

CFD++
FLOW-3D
NASTRAN

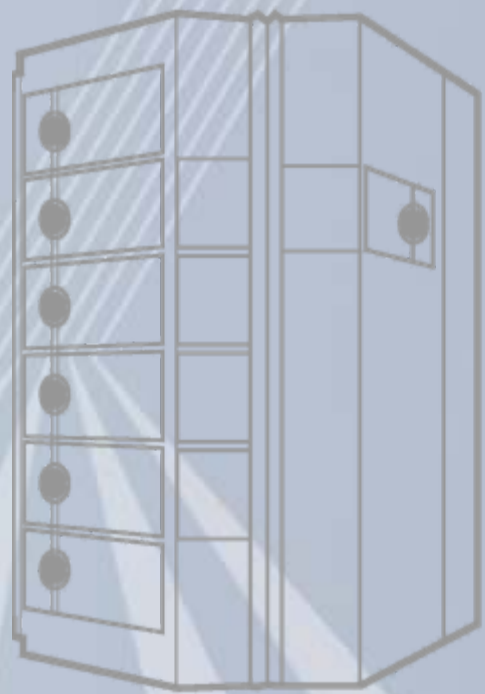
Integrated Vector/Scalar Computing

- Single point of login with complete Linux software environment
- Single Lustre environment provides shared files
- Common ALPS for scheduling mixed applications on both types of compute nodes



BlackWidow

- System Update
- Performance Update
- Programming Environment
- Scalar & Vector Computing
- Roadmap



The Cray Roadmap

Realizing Our Adaptive Supercomputing Vision