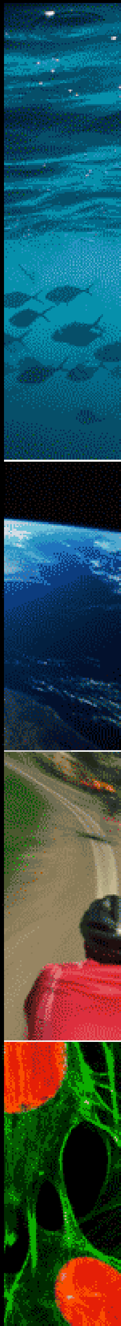
CRAY The Supercomputer Company

The BlackWidow System

Brick Stephenson

CUG 2006



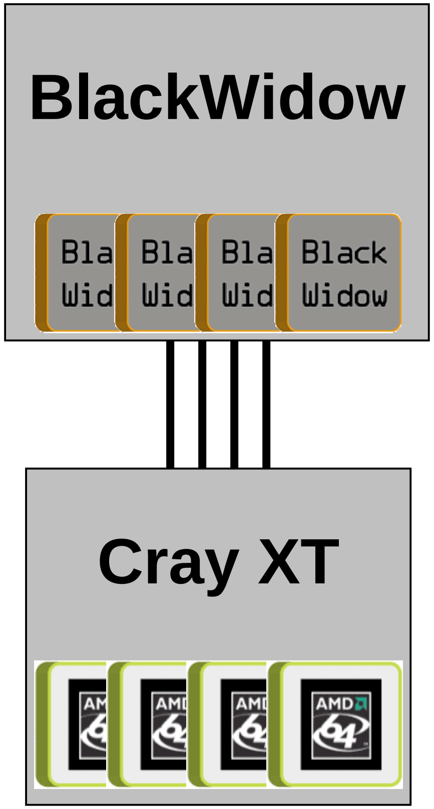
CRAY USER GROUP INCORPORATED

Outline

- System Overview
- Processor Improvements
  - Instruction Set
  - Vector
  - Scalar
- Node Architecture
- Network Interconnect
- Reliability & Scaling Features
- Packaging

BlackWidow System Overview

- Vector compute
- Linux OS



BlackWidow

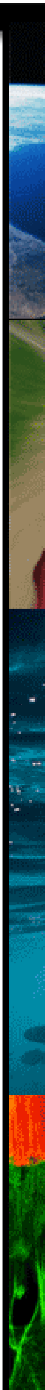
Cray XT

- Services and I/O for BlackWidow system
- Linux OS
- (Optional) Scalar MPP compute

Second generation scalable vector system

BlackWidow CPU

- Cray-developed custom vector processor
- Faster clocks, longer vectors, more pipes
- Improved scalar performance in numerous dimensions relative to Cray X1 and Cray X1E CPUs
- Based on Cray X1 instruction set, with some enhancements



Instruction Set & Vector Changes

- Based on the Cray X1 (“NV-1”) instruction set
- New instructions:
  - Inclusive-OR version of the bit matrix multiply (“Bit matrix compare”)
  - Vector atomic-Ops
    - Fetch{Add,And,Or,Xor}
    - Atomic{Add,And,Or,Xor}
  - Versions of gather and scatter with Sword indices
  - Immediate logicals, integer multiply, and conditional move

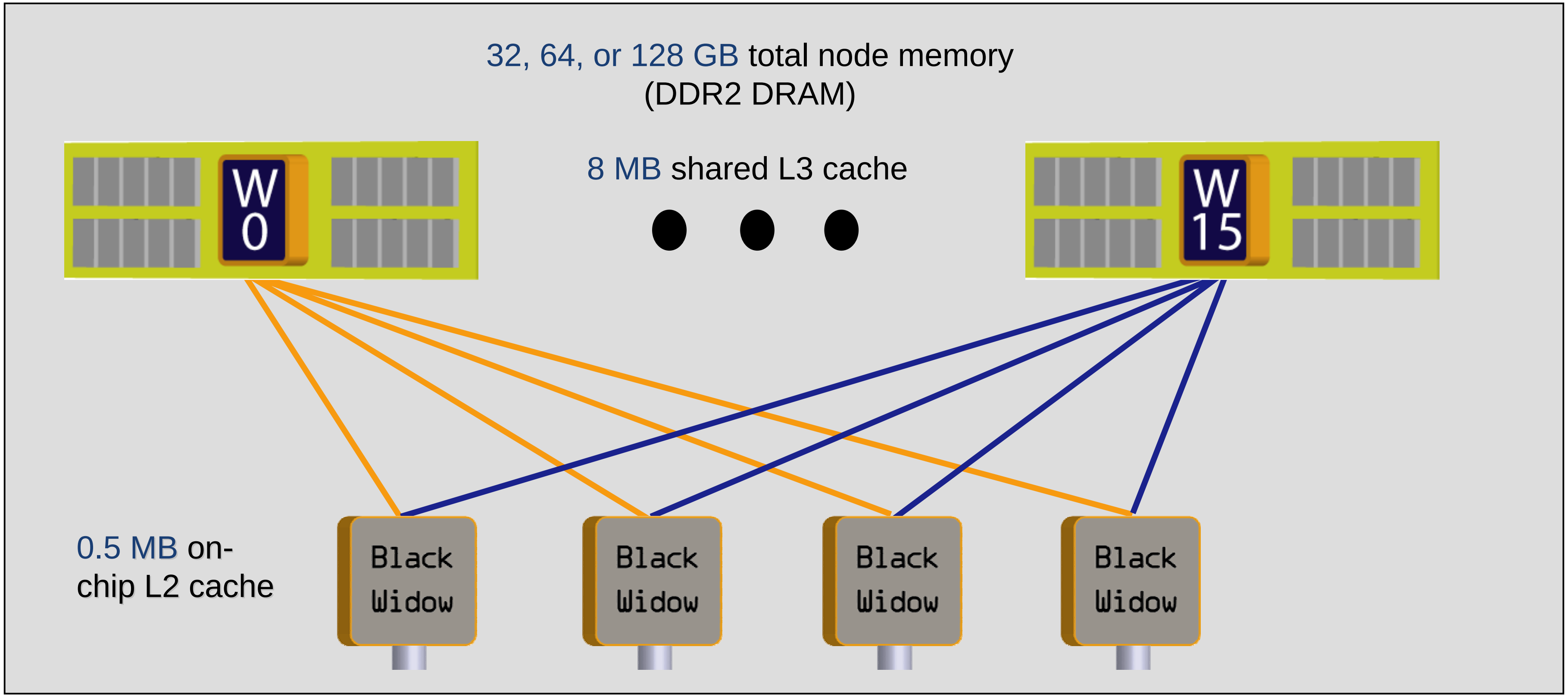
- Maximum vector length increased
- Vector masks increased
- Removed the mod-32 register usage restriction
- Full speed bit matrix multiply

Scalar Improvements

- 4-way instruction dispatch
- Active instruction window enlarged
- Speculative Scalar loads
- Number of outstanding branches increased
- D-Cache hit time reduced
- D-Cache protected from vector traffic
- Level-2 cache hit time reduced
- Local Memory latency reduced

Cray BlackWidow Node

- Globally addressable memory with 4-way SMP nodes
- Two SMP nodes per BlackWidow compute blade



32, 64, or 128 GB total node memory
(DDR2 DRAM)

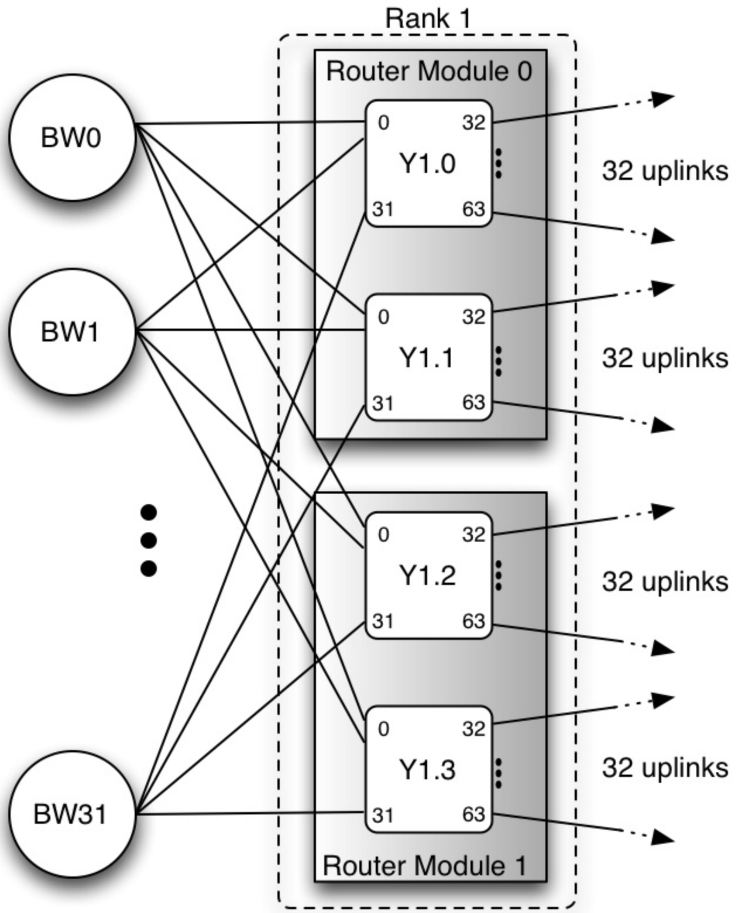
8 MB shared L3 cache

0.5 MB on-chip L2 cache

High Radix Fat Tree Network

Network Topology and Packaging

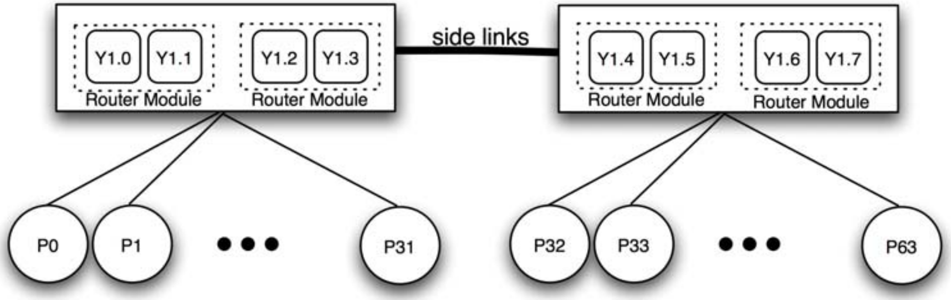
- The BW network is built from YARC high-radix routers
  - 64 ports
- Each BW processor has four network injection ports
  - Each port connects to an independent *slice* of the network
- Scales up to 32k network endpoints (processors)
  - variant of the folded-Clos
- Packaged in:
  - Compute blades, rank1 router modules, and rank2/3 router modules



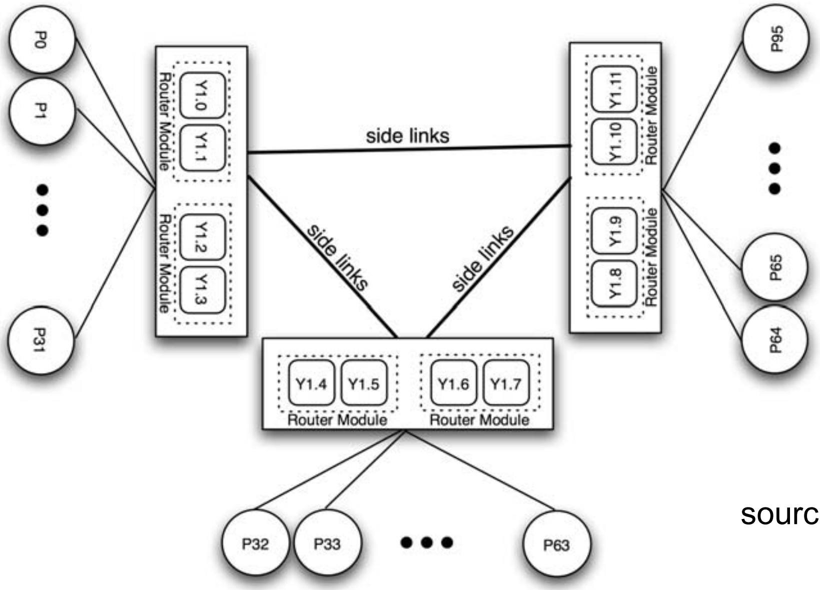
source: Scott, Abts, Kim, Dally ISCA 2006

Topology and Packaging

- Building block is a 32-processor rank 1 sub-tree
- Rank 1.5 network with 64 processors



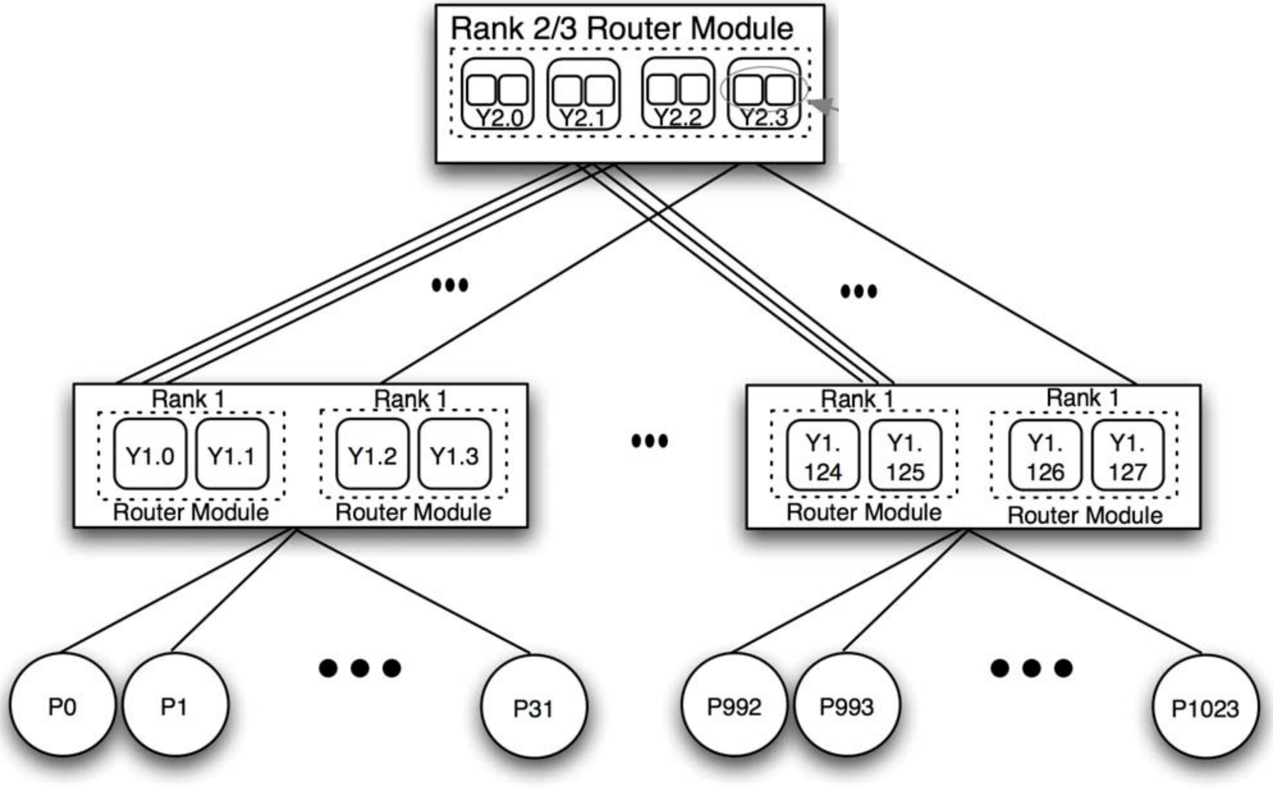
- Rank 1.5 network with 96 processors



source: Scott, Abts, Kim, Dally ISCA 2006

Topology and Packaging

- Rank 2/3 router modules are packaged in a self-contained cabinet
- Each Rank 2/3 module has 32 connectors carrying 256 network links
- Allows a 1024 endpoint Rank 2 network



source: Scott, Abts, Kim, Dally ISCA 2006

Reliability and Scaling Features

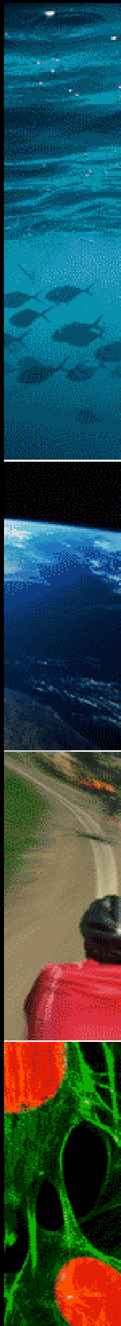
- Fault detection, diagnoses and recovery
  - Hardware Supervisory System (HSS)
  - Comprehensive error detection and logging
  - Timeouts and self-cleansing data paths (no cascading errors)
- Hardware firewalls for fault containment
  - Secure, hierarchical boundaries between kernel groups
  - Protects the rest of the system even if a kernel is corrupted
- Graceful network degradation
  - CRC protection and retransmission to tolerate transient failures
  - Auto-degrade links to tolerate hard failures
  - Hot swappable boards and reconfigurable routing tables
- Full node translation capability
  - Allows scheduling of parallel jobs across an arbitrary collection of processors, with efficient, scalable address translation

  ⇒ Much higher system utilization under heavy workloads

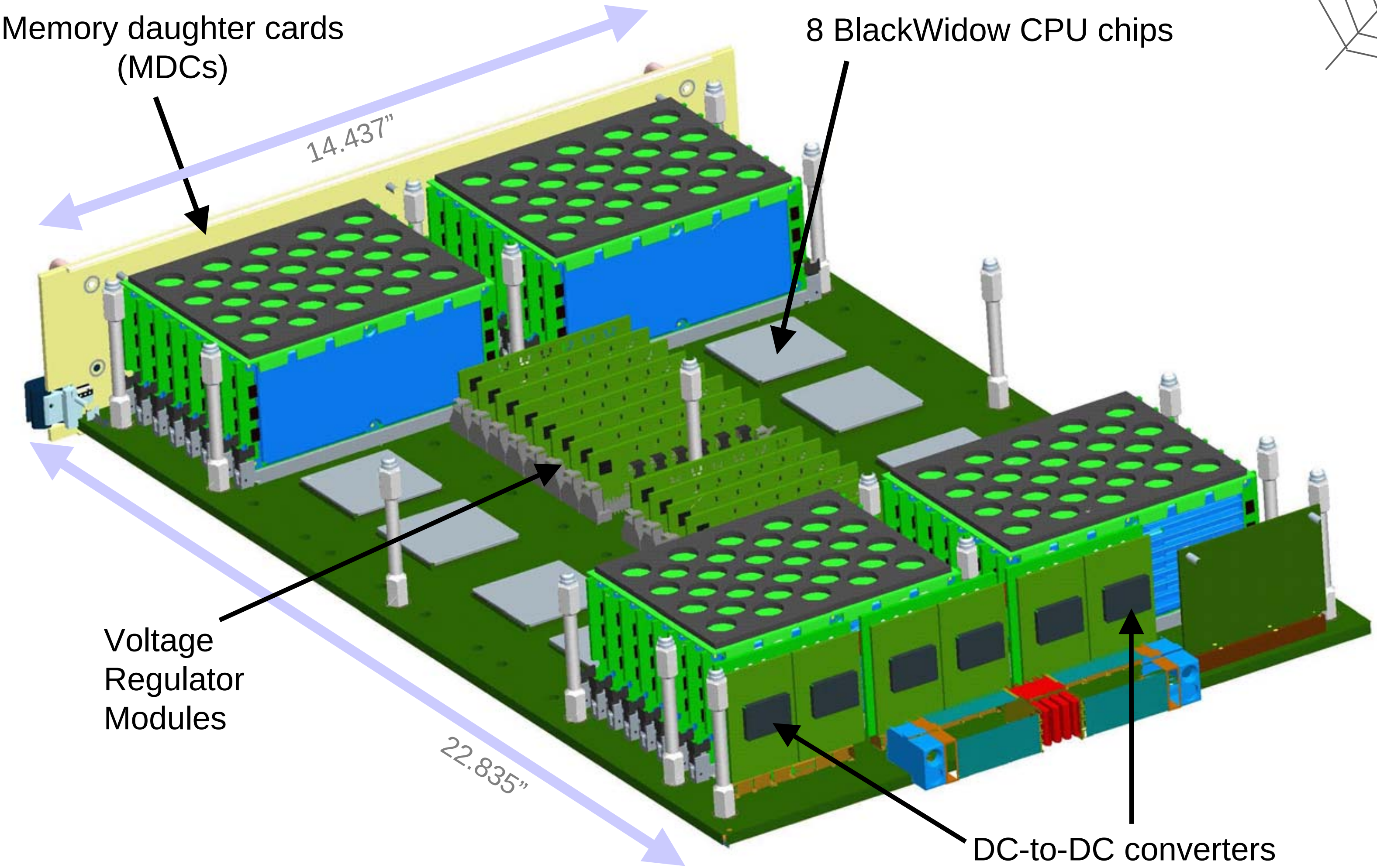
BlackWidow

Packaging

CUG 2006



8 Processor Compute Blade Layout



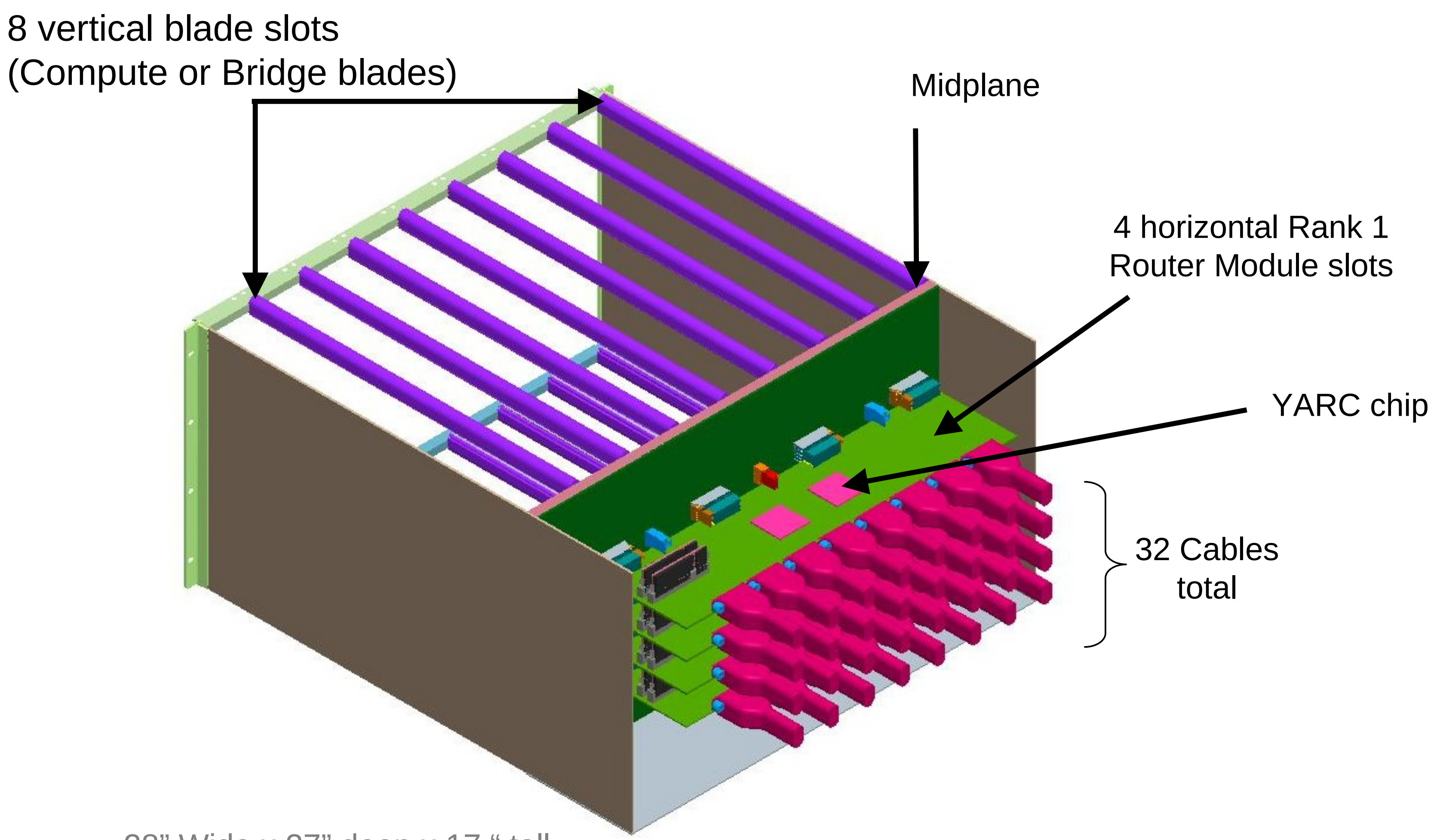
Memory daughter cards (MDCs)

8 BlackWidow CPU chips

Voltage Regulator Modules

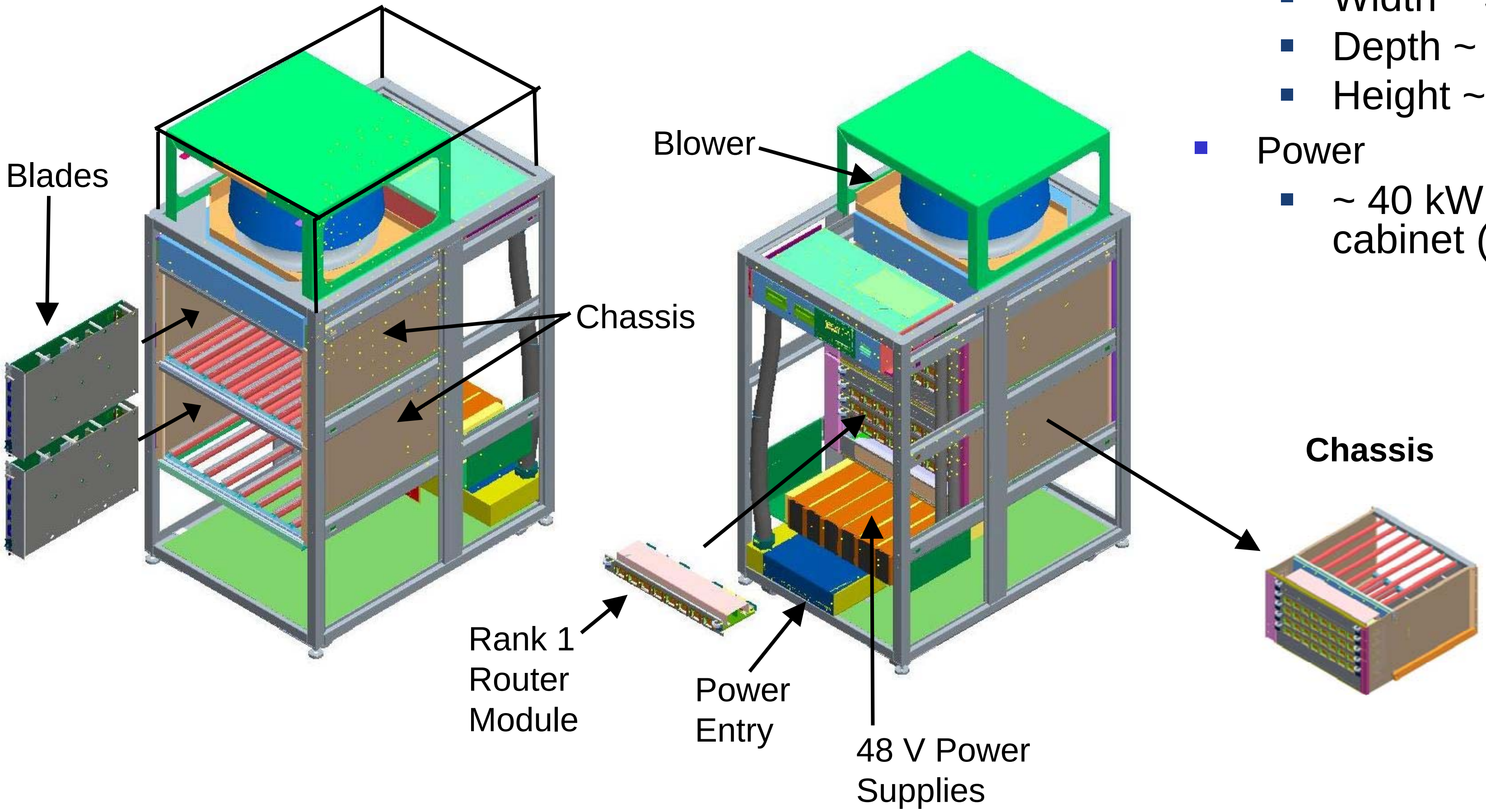
DC-to-DC converters

Chassis/Backplane – Rear View



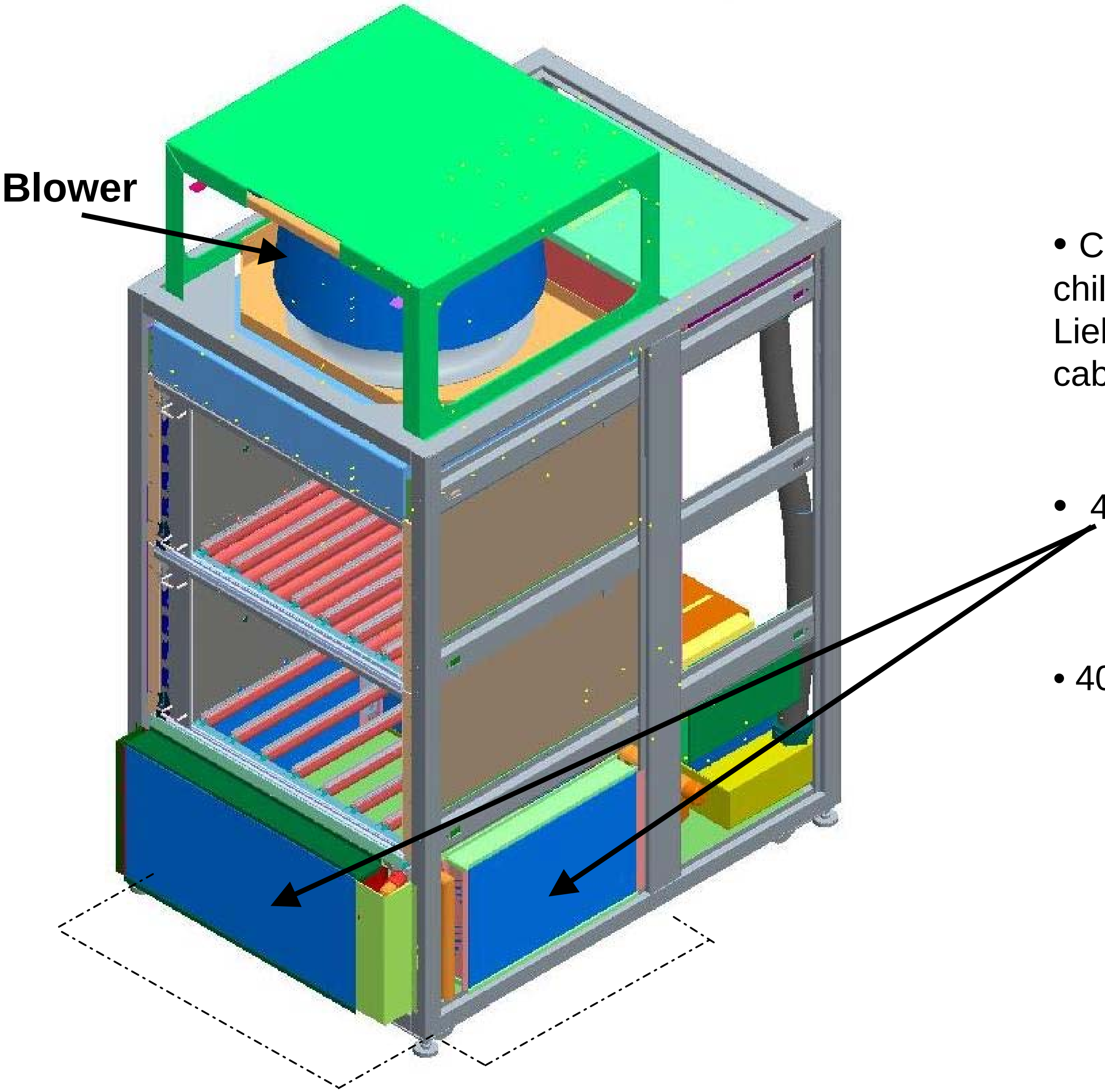
~28” Wide x 27” deep x 17 “ tall

Compute Cabinet – Air Cooled



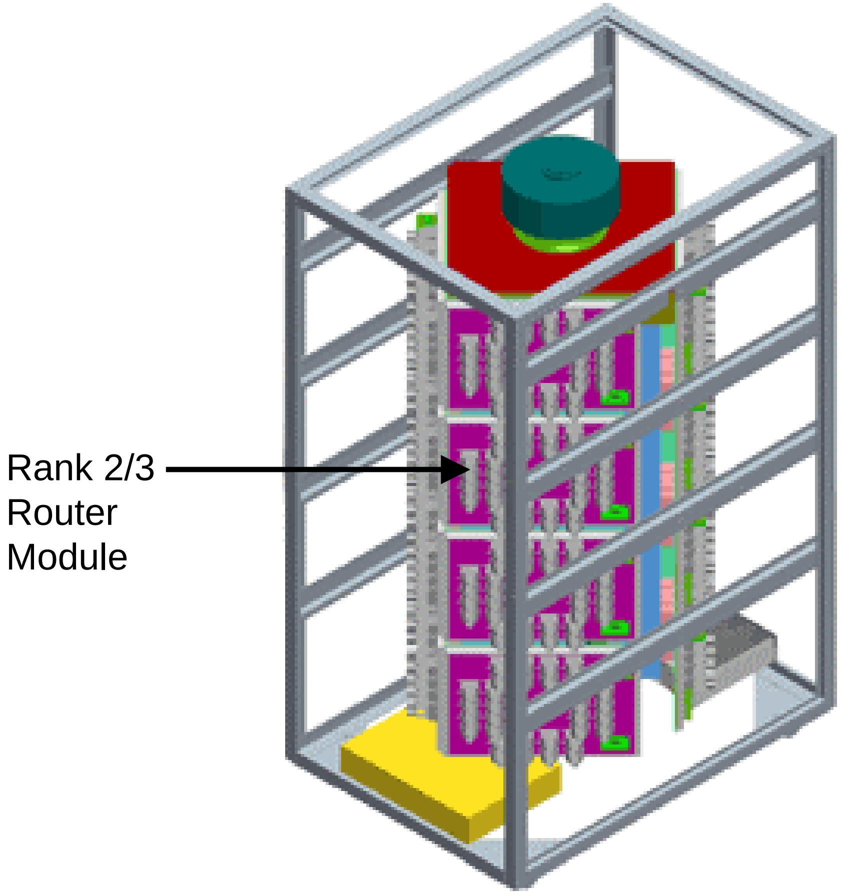
- Dimensions
  - Width ~ 34”
  - Depth ~ 48”
  - Height ~ 78”
- Power
  - ~ 40 kW per cabinet (128p)

Compute Cabinet - Liquid Cooling Option



- Closed loop air re-circulating through chilled-water coils (as if embedding site Lieberts/Pomonas into the Compute cabinet)
- 4 coils located in cabinet base
- 40kW: ~40gpm @ 60F chilled water

R2/R3 Router Cabinet



16 rank 2/3 router modules – 4 for each direction (N,E,S,W)

512 total cables