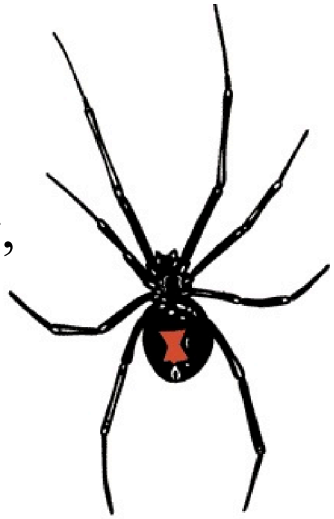
The Cray BlackWidow: A Highly-Scalable Vector Multiprocessor

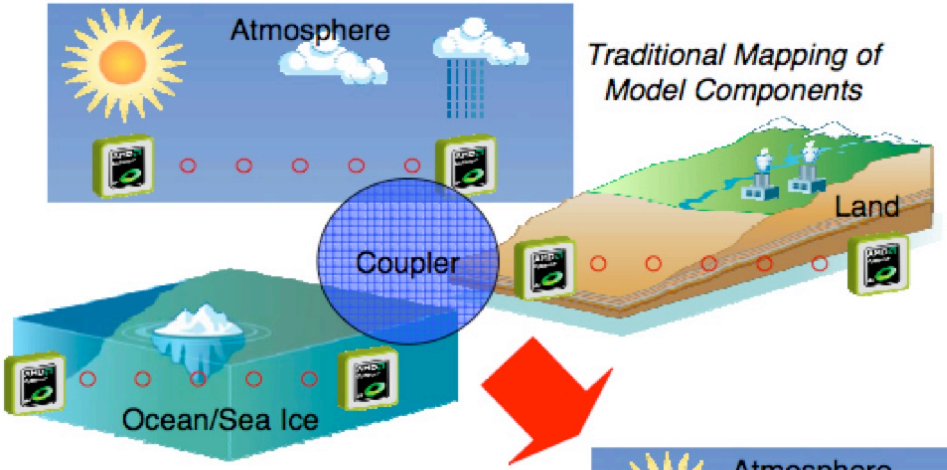
Dennis Abts, Abdulla Bataineh, Steve Scott,
Greg Faanes, Jim Schwarzmeier, Eric Lundberg,
Tim Johnson, Mike Bye, Gerry Schwoerer



Outline

- Introduction
- System Overview
  - node structure
  - addressing model
- Architecture Overview
  - memory system
  - network
- Performance
  - HPCC
  - Overflow
- Summary

Hybrid Mapping of Climate Model Components

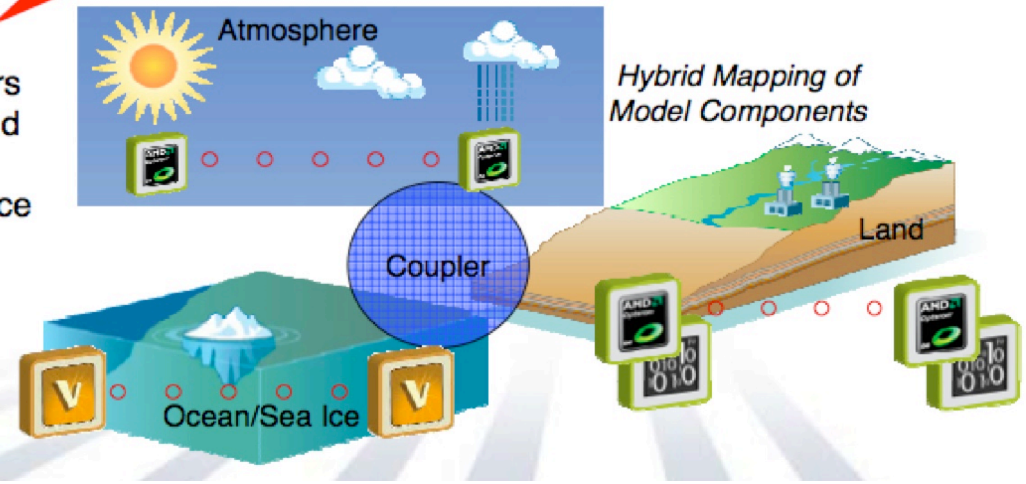


Traditional Mapping of Model Components

Modern climate models are composed of coupled components representing the overall earth system.

Hybrid architectures offer the potential to significantly improve upon the performance offered by uniform processor resources.

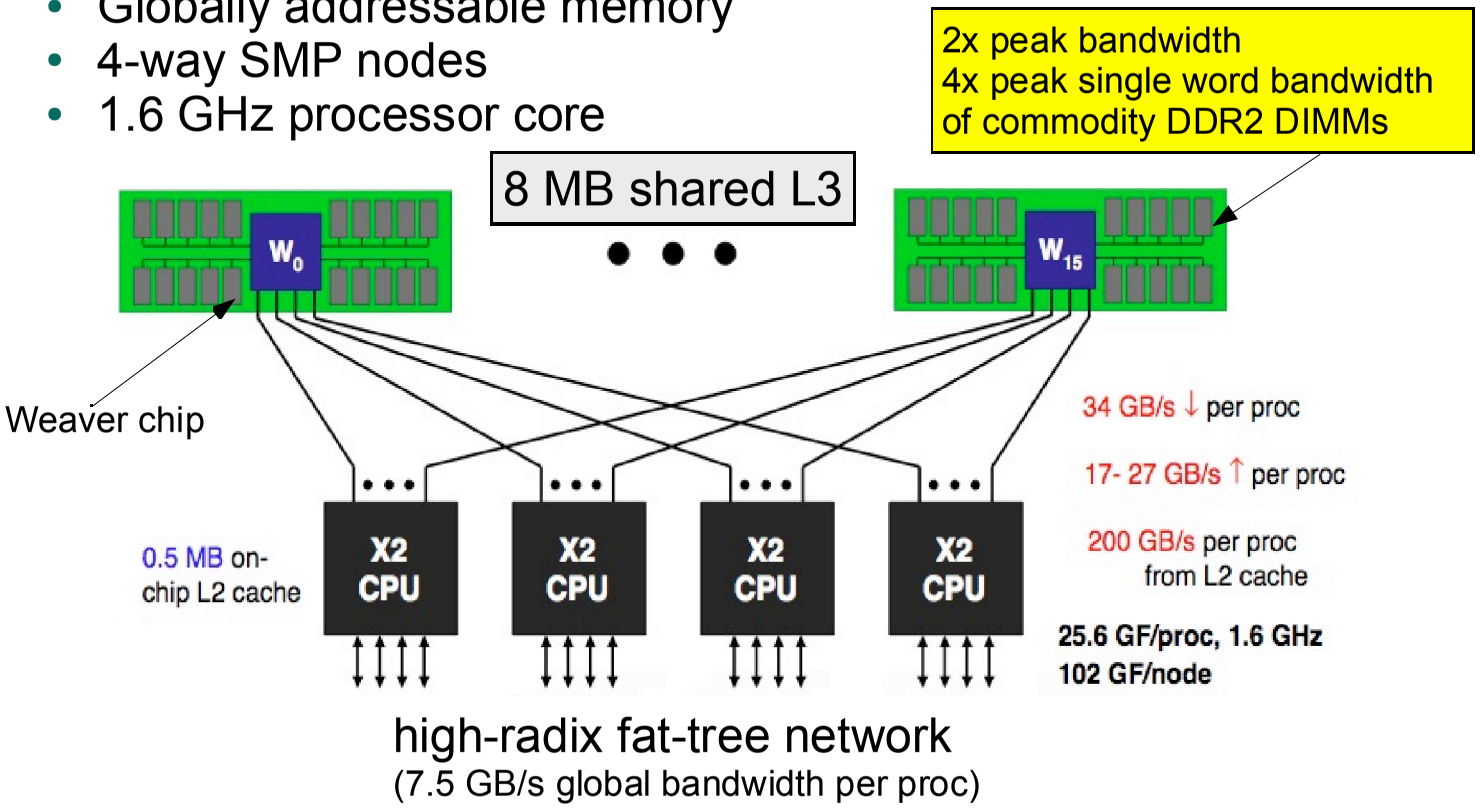
Each processing technology offers it's own unique characteristics and opportunities that can be leveraged to meet the performance demands of each component.



Hybrid Mapping of Model Components

BlackWidow (X2) Node Structure

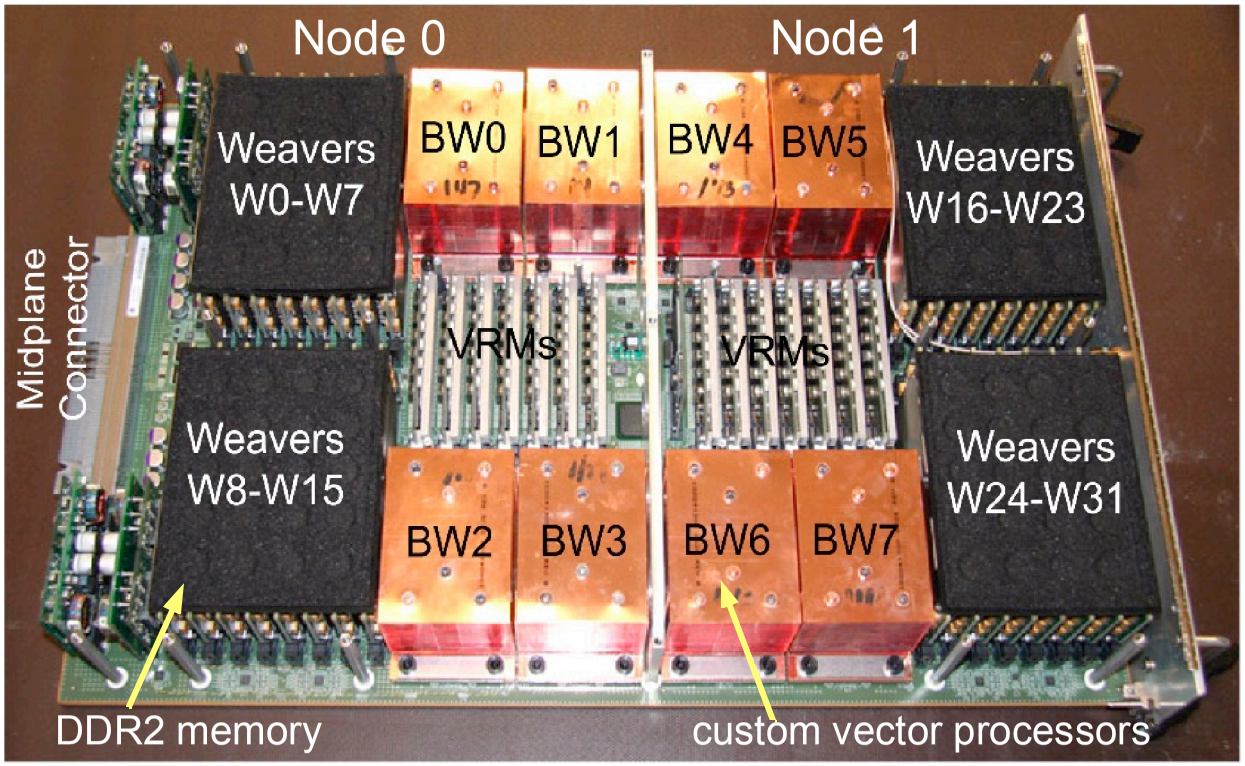
- Globally addressable memory
- 4-way SMP nodes
- 1.6 GHz processor core



high-radix fat-tree network
(7.5 GB/s global bandwidth per proc)

BlackWidow (X2) Compute Blade

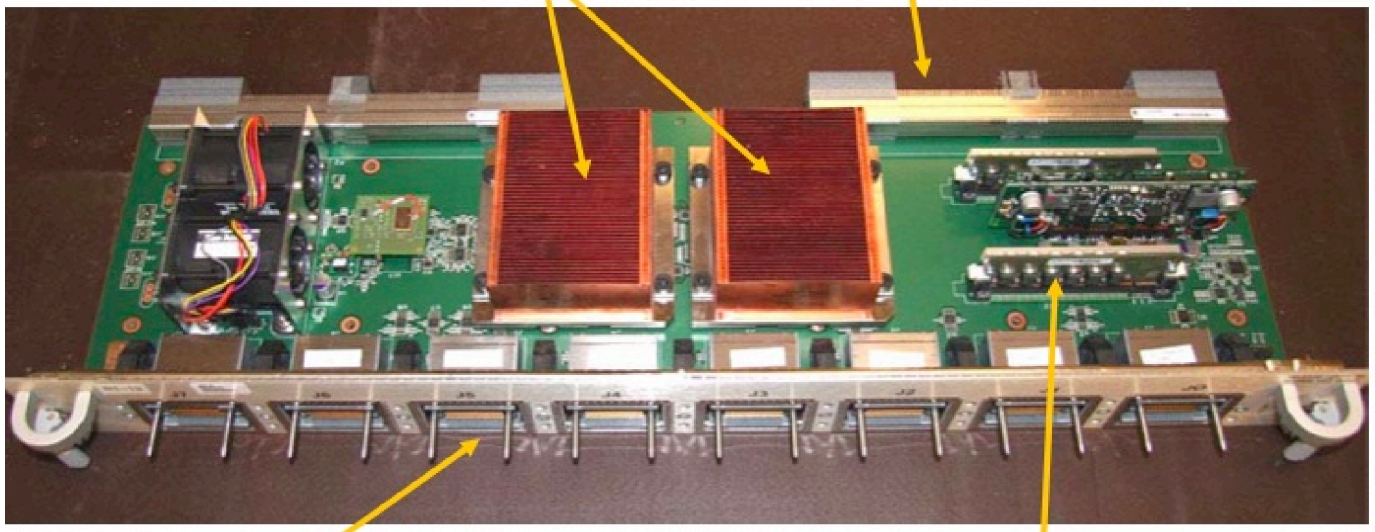
- Two 4-way SMPs per blade



BlackWidow System Architecture

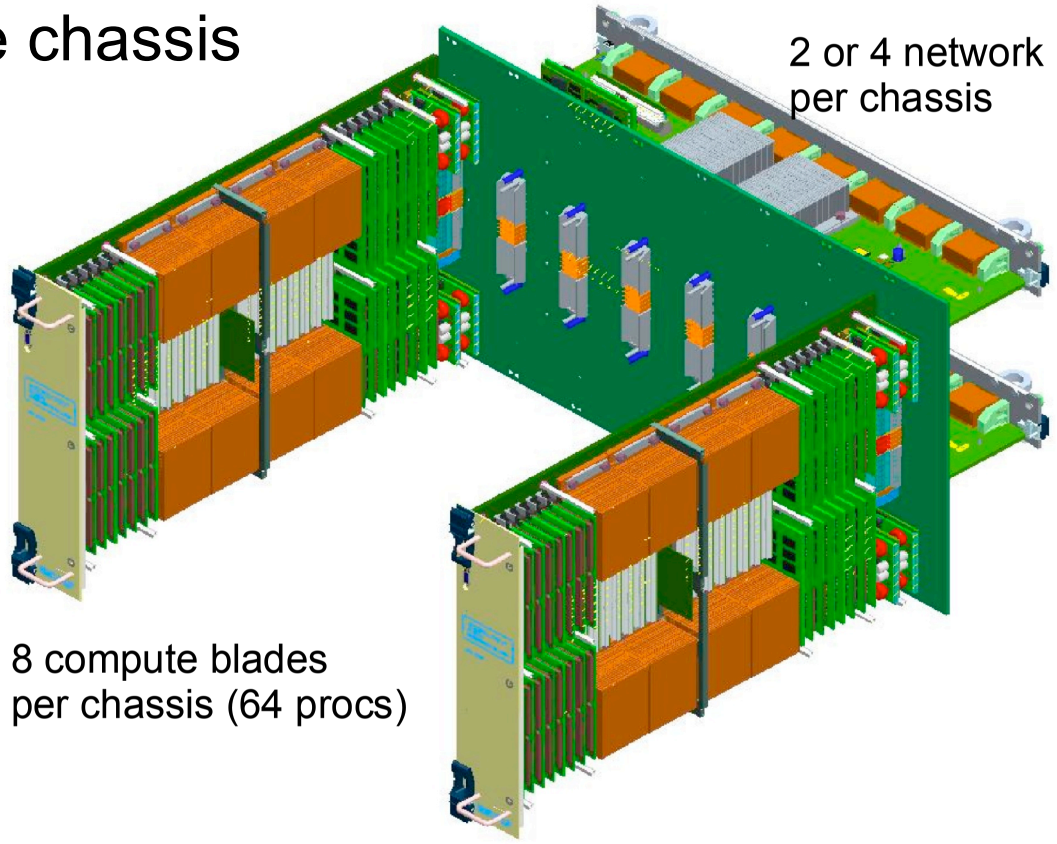
System Packaging

Rank1 Router Card



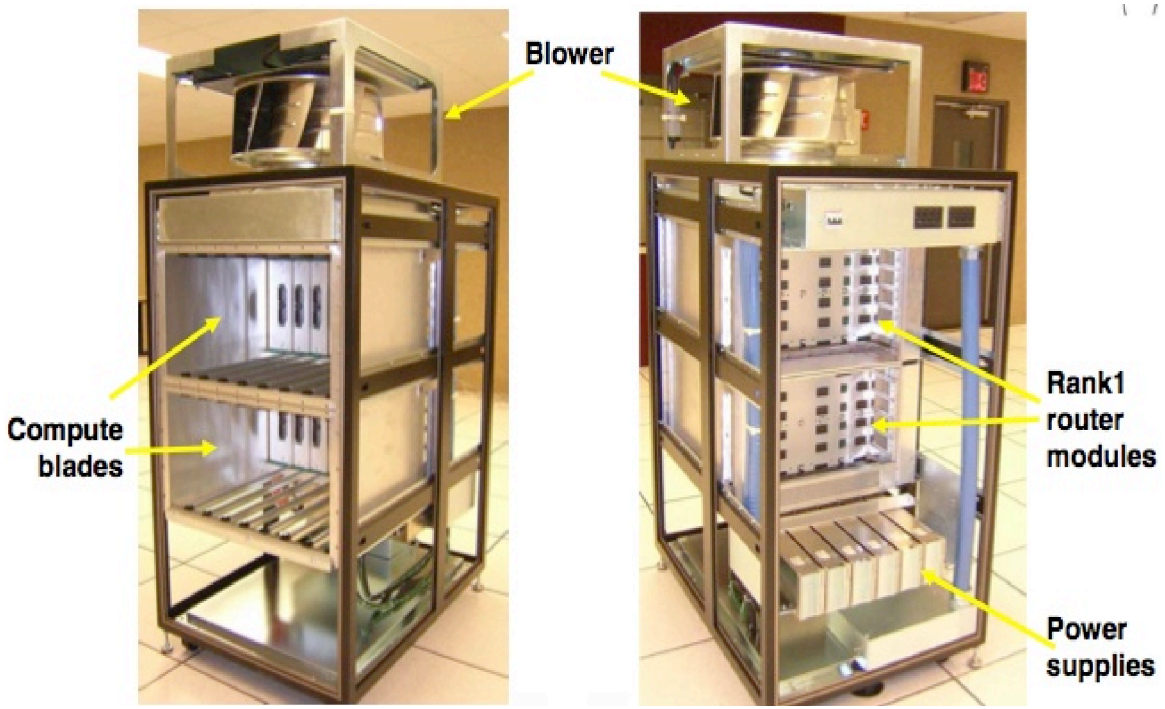
System Packaging

- 8-blade chassis



System Packaging

- Two 8-blade chassis in each cabinet (128 procs per cabinet)

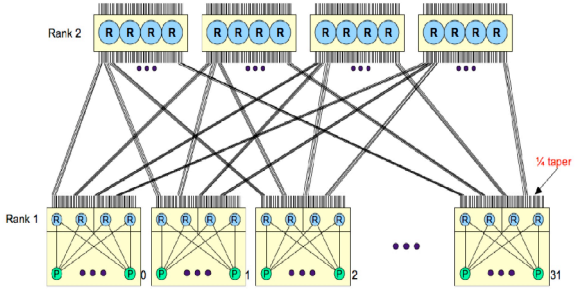


BlackWidow System Architecture

Network Architecture

High-Radix Fat-tree Network

- Key innovation: very high radix router architecture
  - 64 full-duplex ports
  - very low network latency
  - See [Scott, Abts, Kim, Dally – ISCA06]

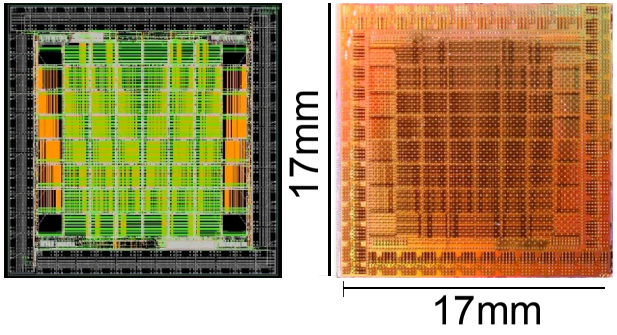


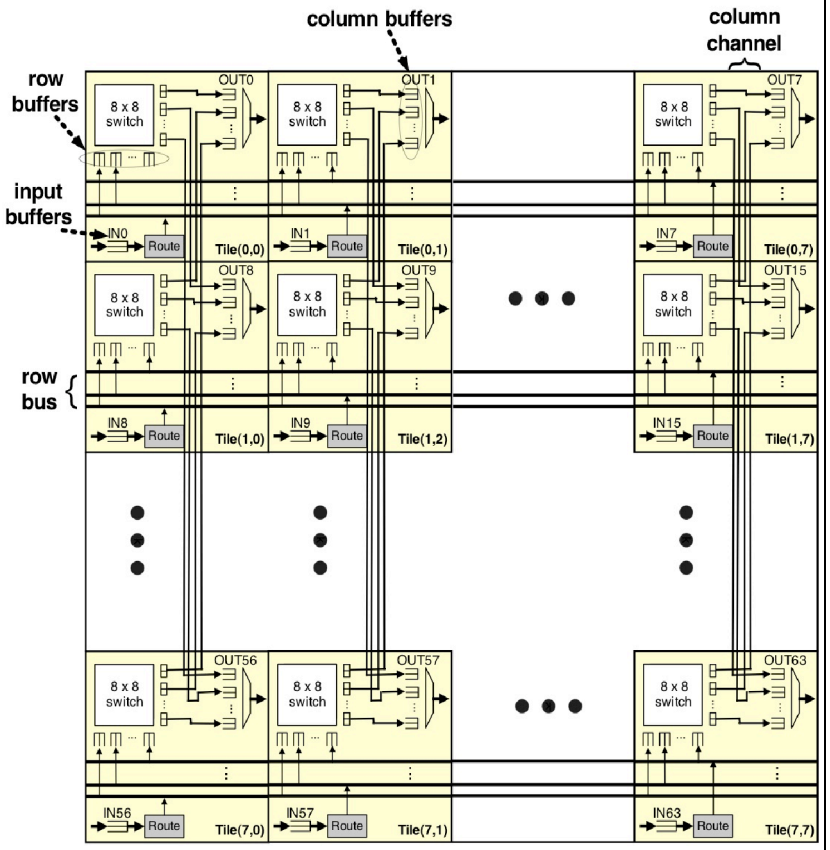
- Performance
  - adaptive routing and bandwidth spreading to reduce network hot-spots
  - can route any permutation conflict free at large scale
  - can configure as much global bandwidth as desired

- Reliability
  - link-level retry on network links
  - CRC protection on internal data paths
  - auto link degrade
  - auto hardware re-routing of adaptive packets
  - flexible routing table configuration to avoid faulty links

Router (YARC) Microarchitecture

- unique tiled router design
- each tile implements a port and 8x8 subswitch
- no global arbitration
- non-blocking with large internal speedup
- routing table per tile for high packet throughput
- 90nm std-cell ASIC @800MHz



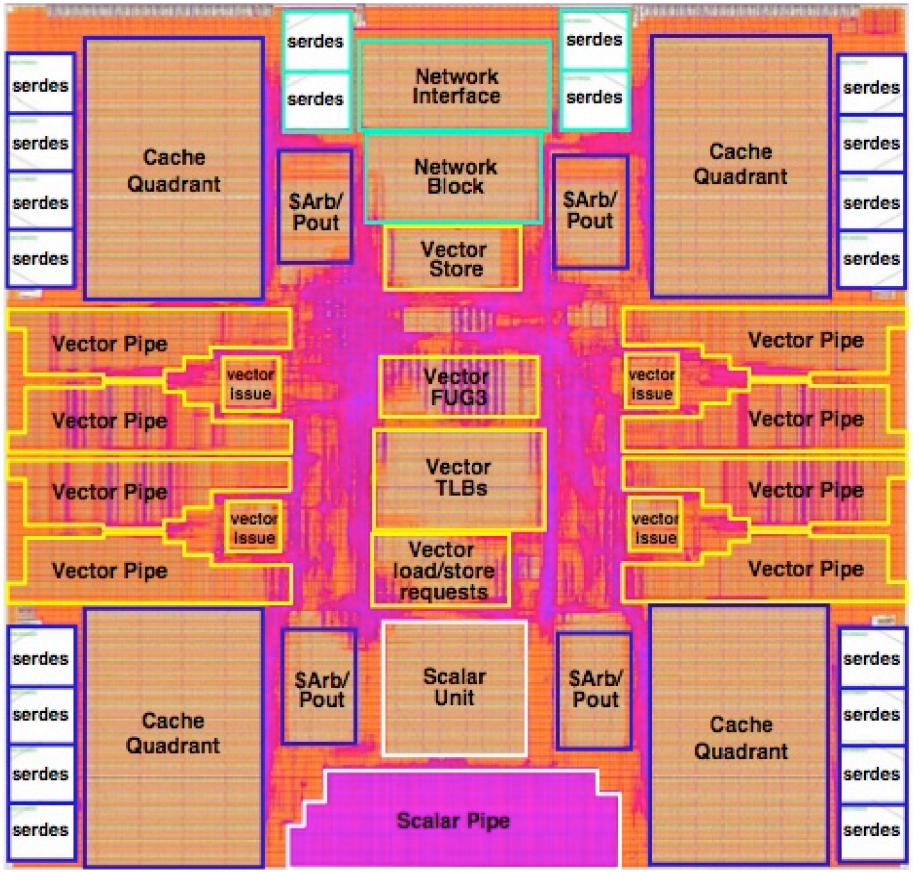


BlackWidow System Architecture

Processor and Memory Architecture

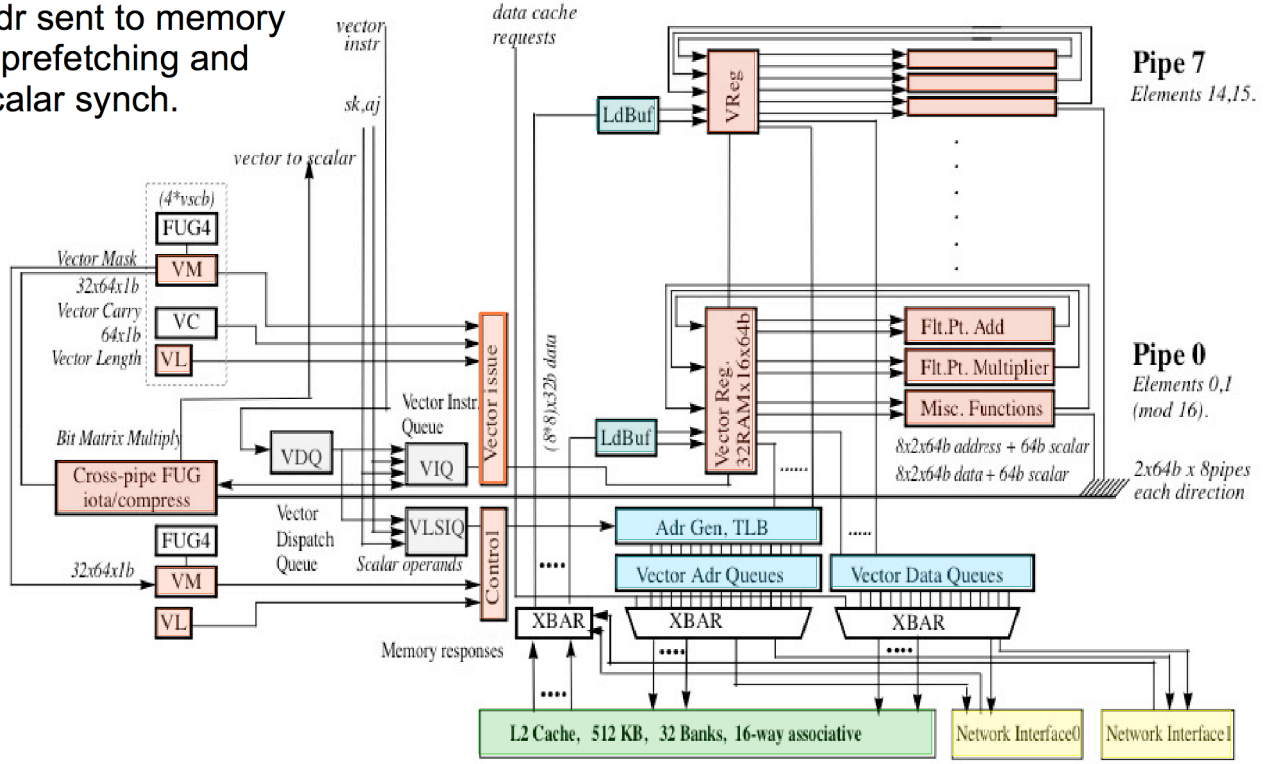
BlackWidow Processor Chip

- 90nm TSMC
- semi-custom
- 16 Proc-to-memory channels
- 4 network channels
- 64-bank L2 cache



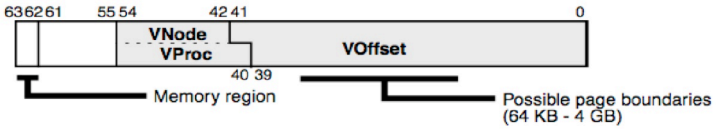
Vector Unit

- Decoupled access/execute unit
- Load buffers allow efficient unrolling; renaming only load targets (up to 4K outstanding)
- Write addr sent to memory early for prefetching and vector/scalar synch.

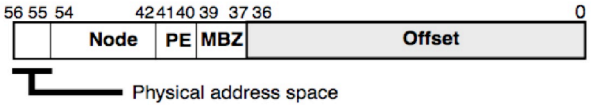


Address Spaces and Translation

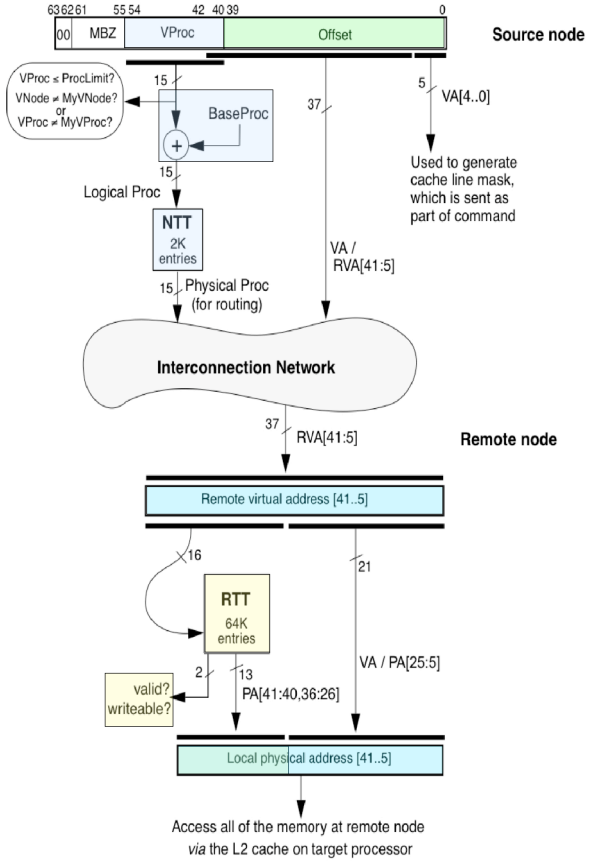
- Native PGAS with node or processor granularities



- 55-bit physical address



- up to 128GB (37 bits) of node offset
- supports both source and remote addr translation
- variable page sizes from 64K to 4GB
- can map the entire address space concurrently



Memory System

- Local memory system design for vector processing
- Reducing latency is critical for scalar cache
- A vector cache is more concerned with reducing the bandwidth demand on local memory
- 32 L2 cache banks support 200 GB/s and over 12 Grefs/s
- Hashing physical address reduces alignment and stride conflicts
- 64 cache coherence directories per node

Coherence Protocol

- Globally cache coherent
- Only cache data local to the node
  - do not cache remote data
  - avoids interventions across the network
  - avoids large fault region of a large SMP
- Strided writes allocate in cache without fill from DRAM
- 32-bit interface to memory (efficient single word operations)

- Only dirty words of the line are written back

- Vector Store Combining Buffer (VSCB) tracks pending store addresses for ordering purposes

- Load “hints” control cache allocation
  - shared, exclusive, or non-allocate

Memory Consistency

- Relaxed memory model
- Vector and scalar references are unordered by default
- Lsync and Gsync primitives provide ordering when necessary
  - Release and Acquire primitives (GsyncR, GsyncA)
- Native scalar and vector atomic memory operations to all of memory

Reliability Features

- Hardware firewalls for fault containment
  - secure, hierarchical boundaries between nodes
  - each node decides who has access to which pages of its memory
  - protects the rest of the system even if a kernel is corrupted
- Graceful network degradation
  - link-level go-back-N retry for transmission errors
  - CRC protection on internal data paths
  - Auto-degrade links to tolerate hard failures
  - Hot swappable boards and reconfigurable routing tables

Reliability Features

- Memory scrubbing
  - hardware scrub engine repairs single-bit errors while it refreshes memory
- Memory request auto-retry
  - tolerates errors on the memory channel
- Spare-bit insertion
  - extra bit can replace a faulty bit in memory
  - hardware histograms support identification of bad bits
- Data poisoning and deferred errors
  - speculative loads that incur an UE will not raise an exception
  - data that is corrupted in the cache is “poisoned”, and the later exception raised iff consumed

BlackWidow System Architecture

Performance

System Configuration for HPCC Benchmarks

	HP XC3000	Cray X2
Processor	Intel Woodcrest Core Duo	Cray BlackWidow
Frequency	3.0 GHz	1.6 GHz (1.3 GHz)
Socket peak	24 Gflops	25.6 Gflops (20.8 GF)
Processor count	32-128	16-64
Threads	1 per core	1
Network	Infiniband 4x @ 5.0 Gbps	YARC 12x @ 5.0 Gbps (2.5 Gbps)
MPI	HP MPI 02.02.05.00	Cray MPICH2 1.0.4

HPCC Benchmarks

System	Number of Sockets	EP-STREAM (GB/s)		EP-DGEMM (Gflops)
		System	Triad	
HP XC3000 (Woodcrest)	16	49.15	1.536	9.29
	32	98.56	1.54	9.49
	64	197.6	1.544	9.17
Cray X2	16	393.0	24.56	21.96
	32 (1)	752.3	23.51	18.36
	64 (1)	1619	25.31	17.98
Relative Performance (X2/Woodcrest)	16	8.0	16	2.4
	32	7.6	15	1.9
	64	8.2	16	2.0

(1) uses slower, prototype hardware

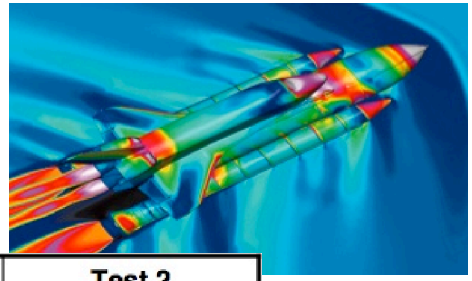
HPCC Benchmarks

System	Number of Sockets	G-PTRANS (Gbytes/s)	G-RA (GUPS)	RandomRing	
				Bandwidth (GB/s)	Latency (μs)
HP XC3000 (Woodcrest)	16	5.142	0.0415	0.300	8.87
	32	8.978	0.0446	0.272	8.99
	64	17.62	0.0402	0.242	9.17
Cray X2	16	28.8	3.01 (2)	2.91	6.75
	32 (1)	24.62	2.75 (2)	1.20	10.43
	64 (1)	67.03	5.41 (2)	1.12	11.57
Relative Performance (X2/Woodcrest)	16	5.6	73	9.7	1.3
	32	2.7	62	4.4	0.86
	64	3.8	135	4.6	0.79

(1) uses slower, prototype hardware (2) uses CAF

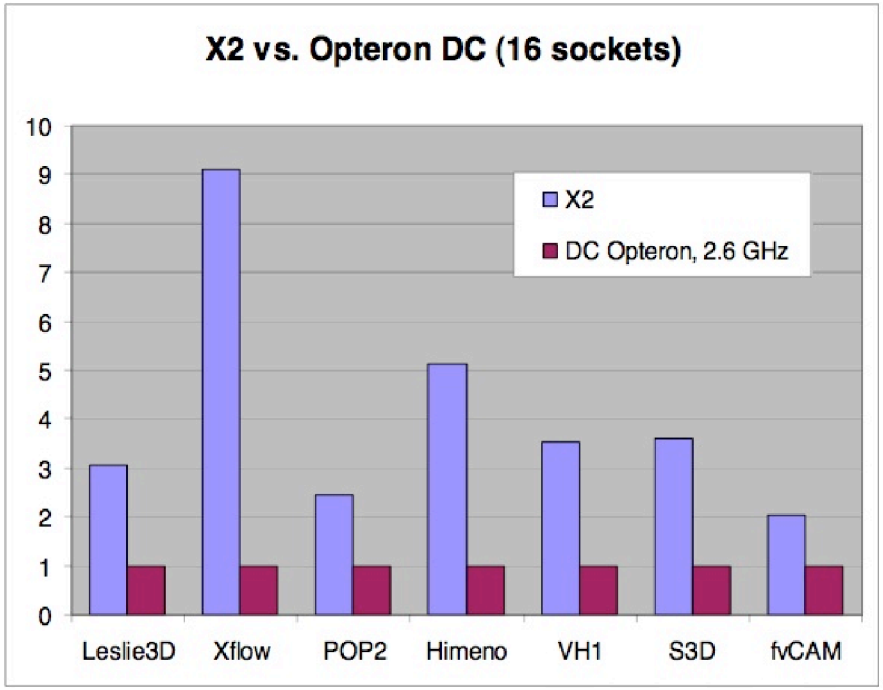
Overflow CFD Benchmark

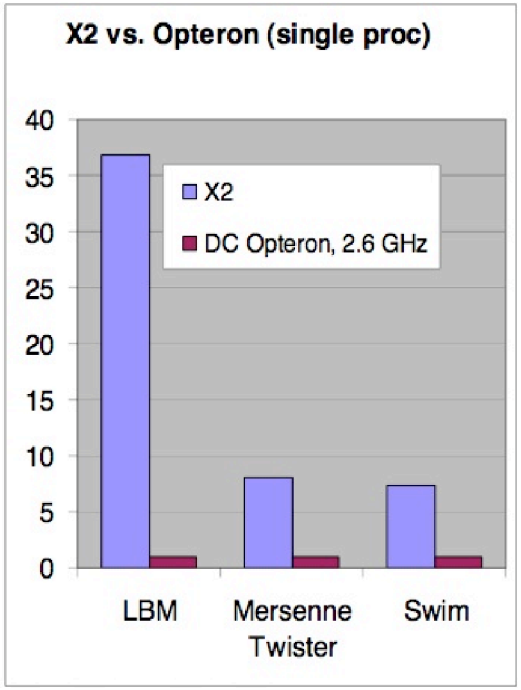
- Limited scalability
- Memory intensive



System	Number of Sockets	Number of MPI processes	Test 1 (sec)	Test 2 (sec)
Clovertown (Quad core, 2.33GHz, 1333 MHz FSB)	1	1	310	3634
	2	4	168	1977
	2	8	107	1124
Cray X2 (1.6 GHz, 5 Gbps ntwk)	1	1	61	494
	2	2	41	295
	4	4	30	195
Relative Performance (X2/Clovertown)	Single processor		5.1	7.4
	Four processors		5.6	10.1
	Two X2 sockets vs. two Clovertown quad-core sockets		2.6	3.8

BlackWidow vs. Opteron





Summary

- BlackWidow is a DSM architecture scalable up to 32K proc
  - decoupled access/execute architecture allows run-ahead execution
  - heterogeneous system with Cray XT5 that combines scalar compute blades (w/ Opteron) for heterogeneous workflows
- Globally coherent, but does not cache remote data
  Relaxed memory model w/ explicit synch primitives
  - release and acquire semantics using GsyncR and GsyncA
  - directory-based hardware coherence engine embedded in the shared L3
- High-radix fat-tree network
  - adaptive routing and bandwidth spreading to balance traffic
  - scales up to 32K processors with a worst-case diameter of only 7 hops
- Robust system design for reliable operation at large scale

BlackWidow System Architecture

Questions?