The Black Widow High Radix Clos Network

S. Scott, D.Abts, J. Kim, and W. Dally
ISCA 2007

http://images.dailytech.com/nimage/1950_cray_small.png

System Overview

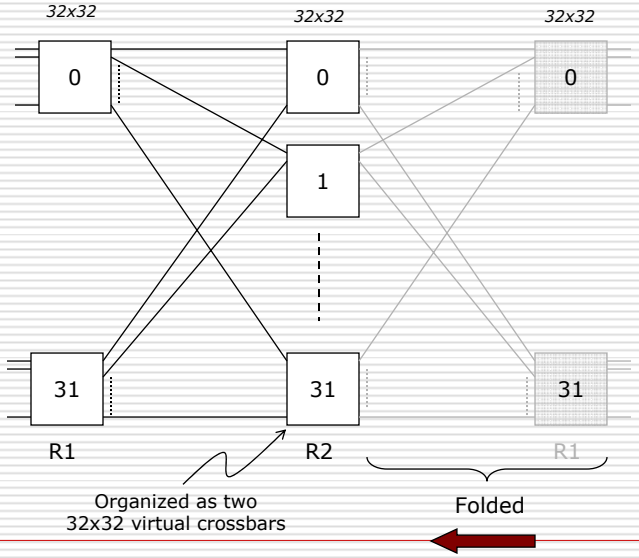
- Designed for communication intensive systems
  - High BW, low latency synchronization
- Shared memory
  - Direct load/store architecture
  - Thousands of outstanding memory references
- 32K processors (72K in principle)
- Folded Clos topology with "side" links
  - Local and global fault tolerant routing
  - Adaptive and deterministic routing
  - High radix and side links

Technology Influence

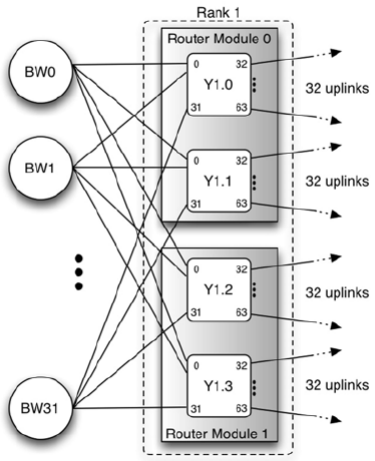
- Pin bandwidth and signaling rates have increased while packet sizes ~ constant
- ASIC technology
  - More logic/pin → more logic/bits/sec
  - Off-chip vs. on-chip signaling rates
  - Wiring vs. buffer tradeoffs
  - Use topology to reduce latency

→ Higher radix topologies

Logical Topology



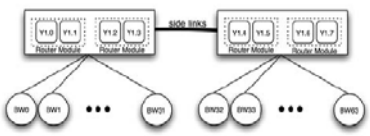
Architecture



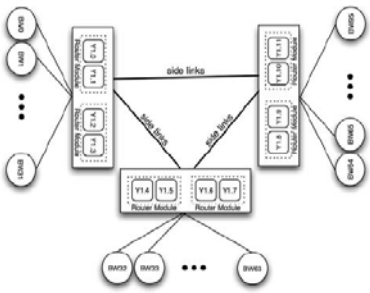
- Each channel is three bits wide
- Radix 64 router chips

From S. Scott, D. Abts, J Kim, and W. Dally, "The Black Widow High Radix Clos Network," *Proceedings of the International Symposium on Computer Architecture*, 2007

Scaling

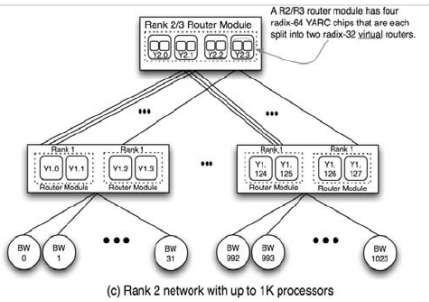


(a) Rank 1.5 network with 64 processors



(b) Rank 1.5 network with 96 processors

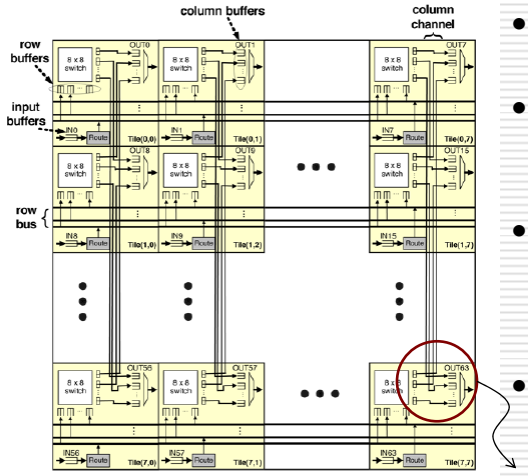
- Use "side" links for incremental scalability
  - Multiple size configurations
  - Half size configurations
  - Physical bandwidth provisioning via packaging



(c) Rank 2 network with up to 1K processors

From S. Scott, D. Abts, J Kim, and W. Dally, "The Black Widow High Radix Clos Network," *Proceedings of the International Symposium on Computer Architecture*, 2007

Follow the Packets



- From input port to column switch
- From column switch to output port multiplexor
- Two stage arbitration
  - For output of tile switch
  - For output port
- Note the size of the arbiters

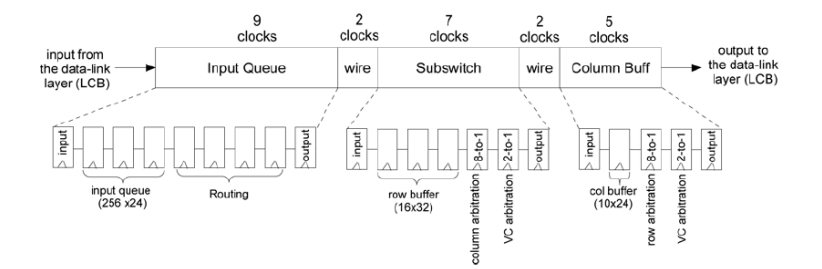
Note wire length due to co-location of output buffers

From S. Scott, D. Abts, J Kim, and W. Dally, "The Black Widow High Radix Clos Network," *Proceedings of the International Symposium on Computer Architecture*, 2007

Design Tradeoffs

- High radix routers
  - Scaling of arbitration costs with input queuing
  - Hierarchical design
  - NRE costs →tiled architecture
- Where do I use my abundance of wires?
  - Bus across the row tiles
  - Point to point across the column tiles

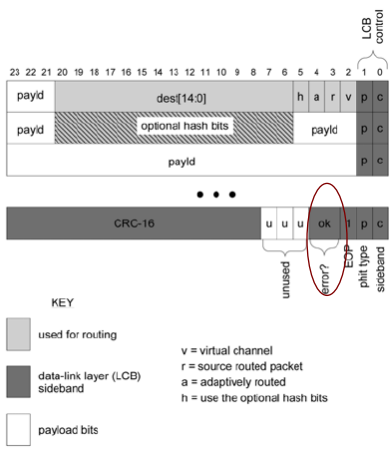
The YARC Pipeline



- 25 stages → no load latency of 31.25 ns
- All major blocks and row and column buses are pipelined
- 24-bit (phit) internal buffer
- VCT flow control externally, wormhole internal
- Variable length packets

From S. Scott, D. Abts, J Kim, and W. Dally, "The Black Widow High Radix Clos Network," *Proceedings of the International Symposium on Computer Architecture*, 2007

Packet formats



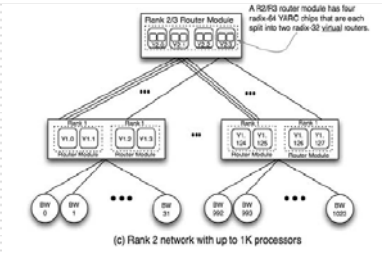
- Source routing for maintenance only
- Optional hash used for deterministic routing
- 2 VCs for request-reply
- 256 phit buffers – link delay sized
- Retransmit handled by credit management
- Soft errors handled by the NI
- Physical layer
  - Mapping 8 lane SERDES macros
  - Channel swizzling

From S. Scott, D. Abts, J Kim, and W. Dally, "The Black Widow High Radix Clos Network," *Proceedings of the International Symposium on Computer Architecture*, 2007

Routing

- Routing – partially adaptive
  - Up routing is adaptive or deterministic
  - Down routing is deterministic
- Requests are ordered
  - Memory consistency model
  - Requests only use deterministic routing
- Destination-based, table driven routing in each tile
- YARC ports are initialized to physical and logical port numbers

Up Routing: Organization



For processors 96-127
Root detect = 0x0060
Mask = 0x001F

- Root detect register for locating the nodes accessible downstream
- Mask for identifying address bits
- Node in the sub-tree if unmasked destination and root detection bits match
- Associative routing table to direct packets
  - Matching entry identifies destination sub-tree
  - May be uplink or side link

From S. Scott, D. Abts, J Kim, and W. Dally, "The Black Widow High Radix Clos Network," *Proceedings of the International Symposium on Computer Architecture*, 2007

Up Routing: Adaptive

- Adaptive if header bit is set
  - Produce a 64 mask of feasible output ports
  - OR operation to produce column mask
- Route to column based on input buffer occupancy
  - Break ties via matrix arbiter in general
  - Heuristic
    - Check MSB of buffer occupancy
    - Round robin arbitration
- Route to column based on occupancy
  - Use the row mask to identify candidates
  - Use 2 bits of buffer occupancy

Up Routing: Deterministic

- Ex-OR of input port and destination
  - Spread the traffic across up links
- For memory addresses optionally include bits of the destination address for further spreading
  - Retains ordering relationship between request to same memory location

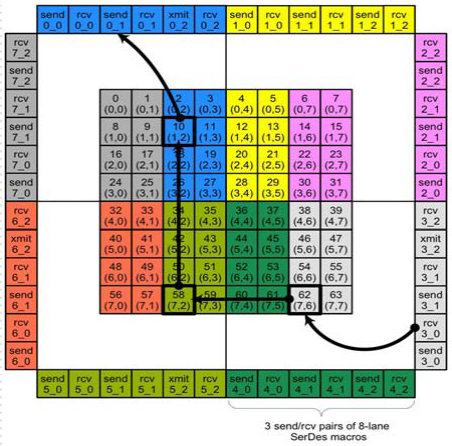
Down Routing

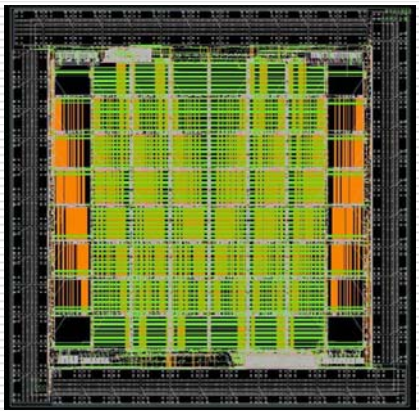
- Deterministic
  - Pick the set of bits in the address
  - Map these to a logical output port
  - Remap table to determine physical output port
- Enables "portability" of YARC chip: use across different nodes in the hierarchy

Fault Tolerance

- Fixed number of retries on transmission failures
  - Note the use of credits to control retransmission
- Error notification for failed links
  - Graceful degradation on channel widths during MTTR
- Soft errors via CRC
  - Error code in header to direct soft error packets to destination NI
- Timeouts on packet injection
- Routing table update to avoid faulty components

Implementation





- 800 MHz core clock, 6.25 Gbps links
- Over half the chip is memory

From S. Scott, D. Abts, J Kim, and W. Dally, "The Black Widow High Radix Clos Network," *Proceedings of the International Symposium on Computer Architecture*, 2007

Comparison to Conventional Wisdom

- Simpler routing than torus
  - No turn or channel restrictions to enforce
  - Addressing is simpler for load balancing
- Lower cost for same bisection bandwidth?
- Easier for partitioning
  - Support for virtualization
- VCT on the links and Wormhole within the switch
  - Cost of on-chip buffers in a tile
- Design space and the degree of the tile switch
  - Range from Xbar to buffered crossbar

Summary

- New design solution for large scale chip-to chip interconnect
- Revisit on-chip networks
  - Abundance of routing resources
  - Scarcity of buffer resources
  - Complexity of the routing function
  - Switching and arbitration does not scale
    - What can fit in a single clock cycle?
  - Locality of design becomes more critical