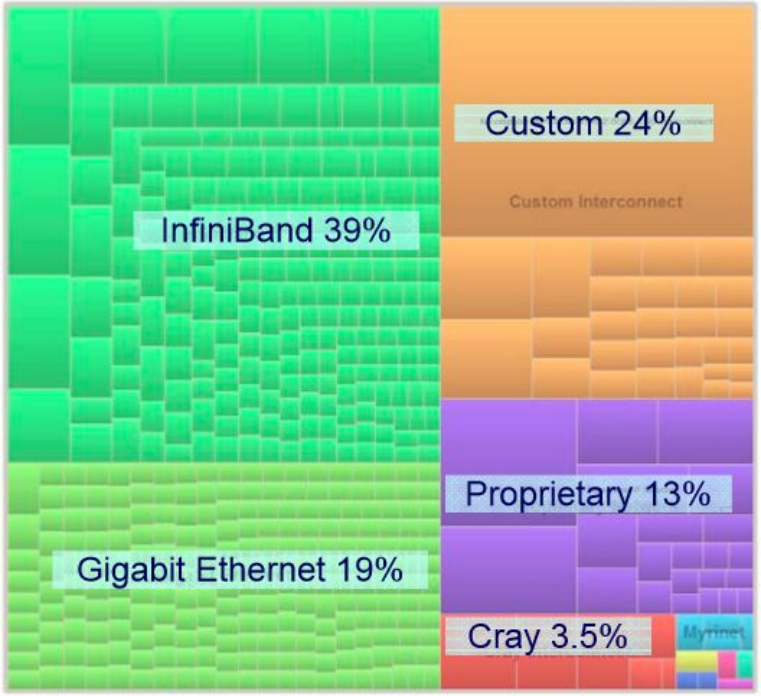
Cray Gemini System Interconnect

Super Computing Interconnects

Scott Cai & Ryan Oberlitner

Overview on Supercomputer Interconnect Family

Interconnect Family Top500 Treemap – Performance (Nov.2011)



Cray Network Evolution

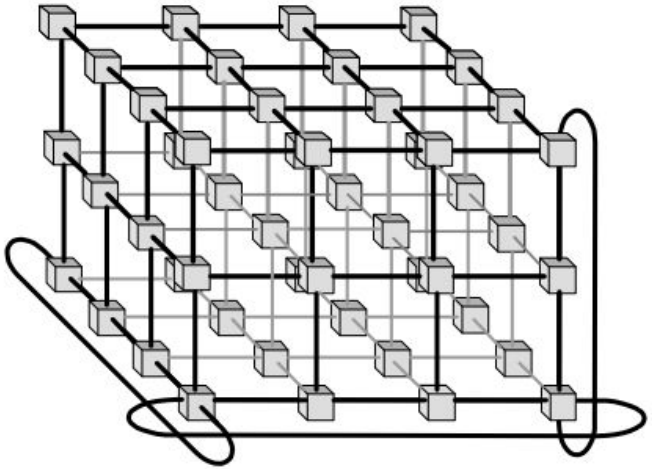
SeaStar
- Built for scalability to 250K+ cores
- Very effective routing and low contention switch

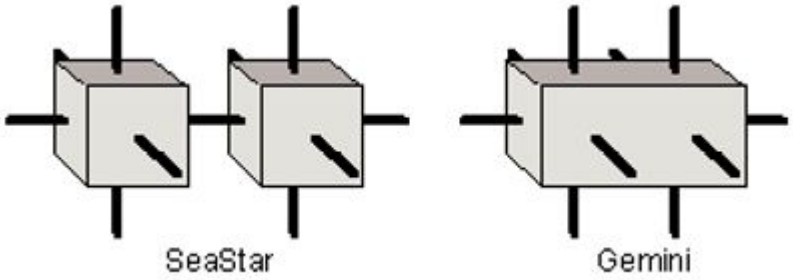
Gemini
- 100x improvement in message throughput
- 3x improvement in latency
- PGAS Support, Global Address Space
- Scalability to 1M+ cores

Aries
- Cray "Cascade" Systems
- Funded through DARPA program
- Details not yet publically available

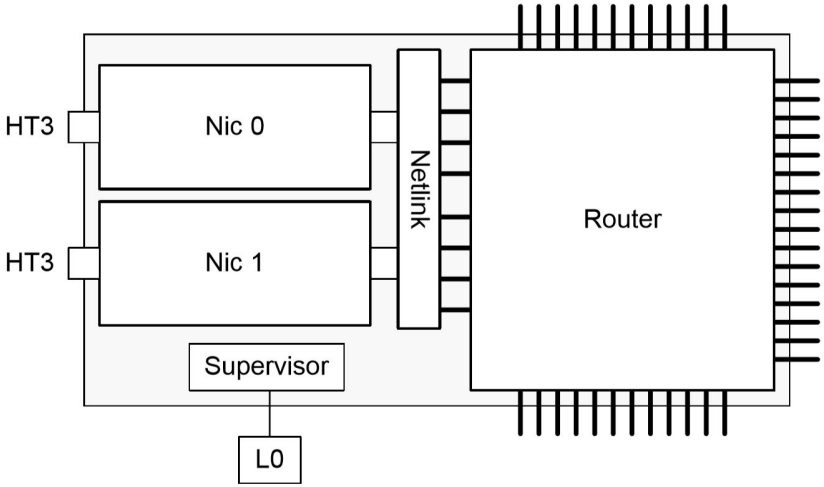
The Gemini system

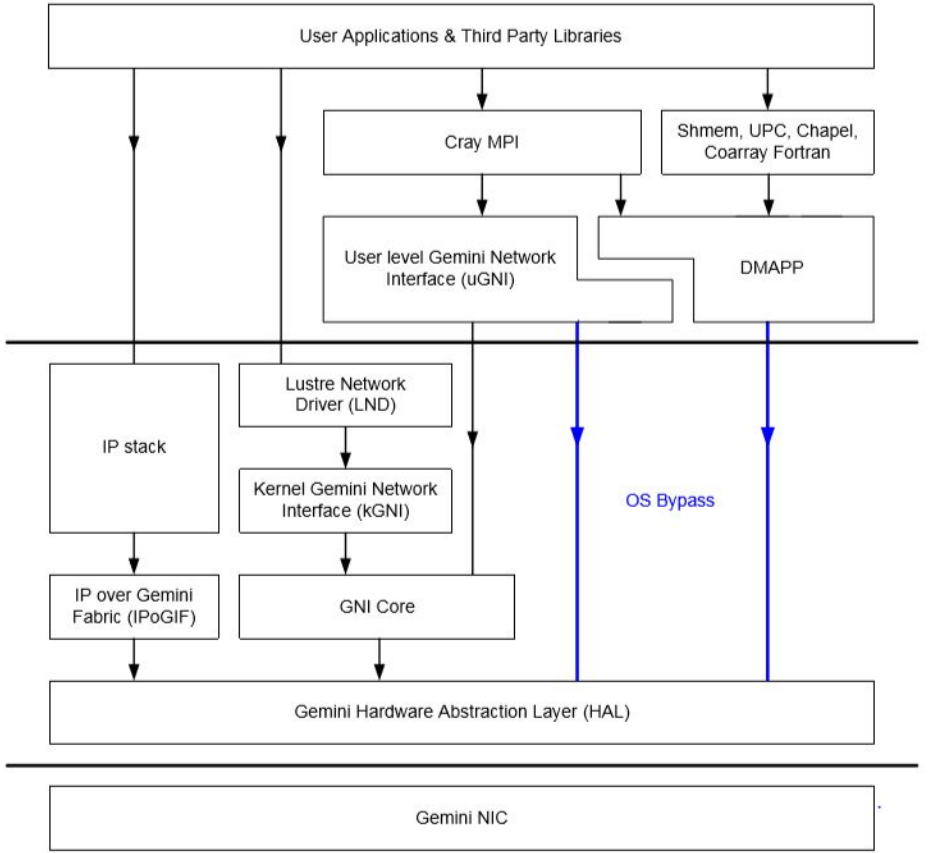
- MPI Network used by Cray's supercomputers
- Uses 3D torus network topology
- CRC checked at each node
- Each physical chip has 2 nodes
- Sliding window protocol
- Data transfers bypass OS





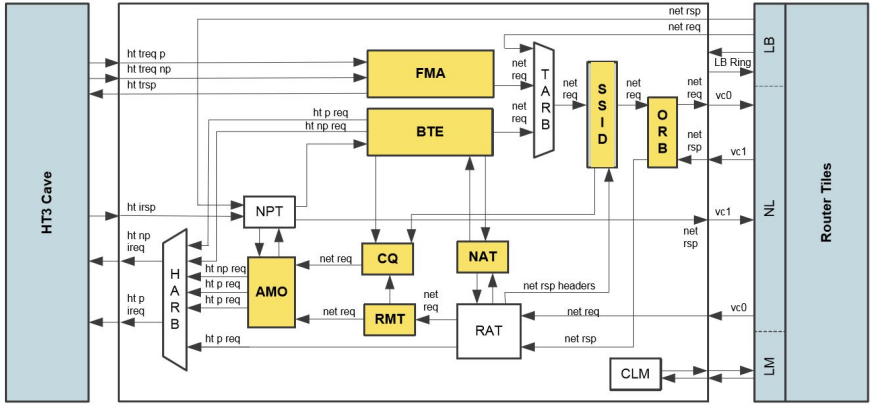
Gemini Software Stack





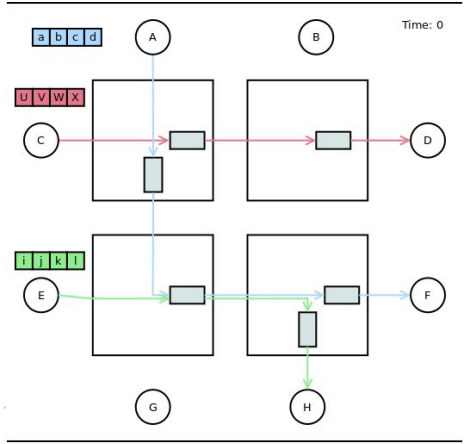
Gemini Network Interface Controller (NIC)

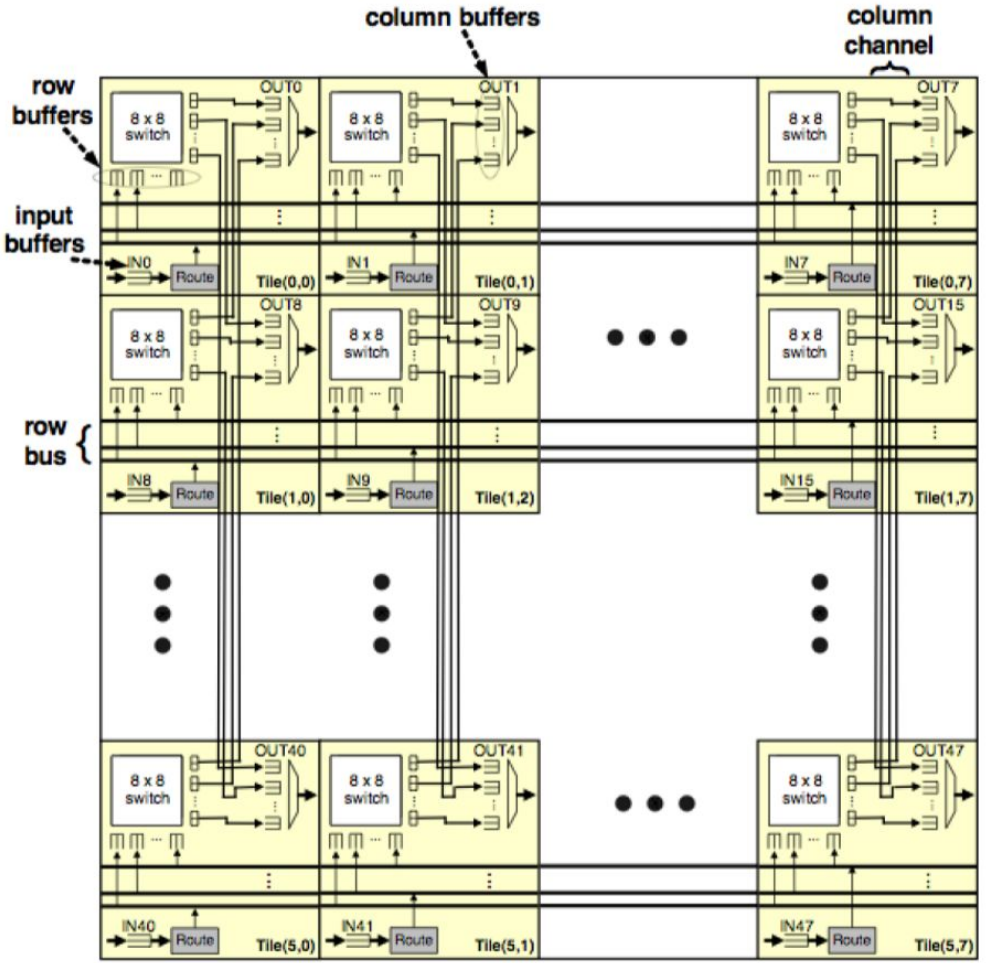
- Atomic Memory Operations (AMO)
  - Get, put, etc.
- Fast Memory Access (FMA)
  - 64 byte remote read/writes
  - Low latency
- Block Transfer Engine (BTE)
  - Larger transfers between local and remote memory
  - Up to 4GB without CPU involvement
- Completion Queue (CQ)
  - Notifications of FMA or BTE
- Synchronization Sequence Identifier (SSID)
  - Packets need not arrive in order, nor be reordered
  - Receive Message Table (RMT) tracks received packets



Gemini Router

- Tile: 1 input port, 1 output port, 1 8x8 switch, buffers
- Virtual cut-through flow control
- Wormhole switching

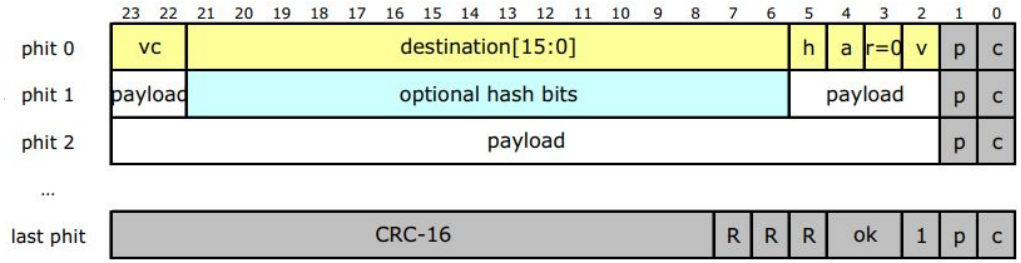




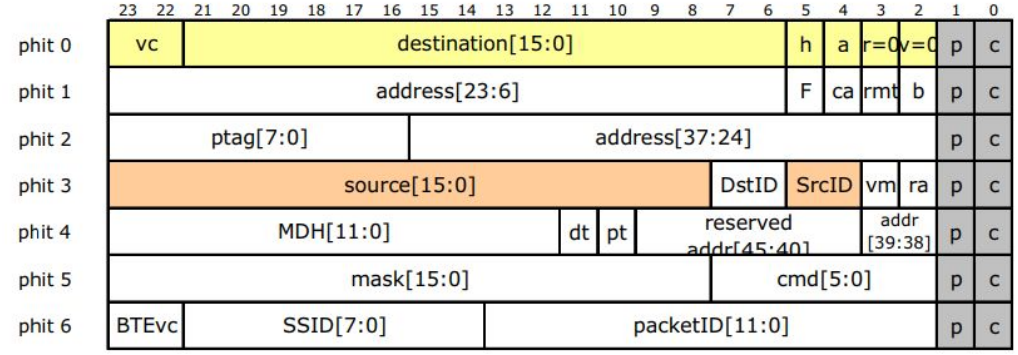
Packet Format

- Uses 24-bit ‘physical units’ (phits)
- Last phit signals end of packet
- 18-bit node addresses
- Up to 45-bit virtual memory address

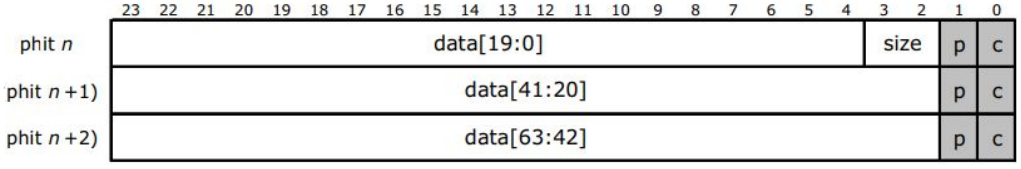
General Network Packet Format



Network Request Packet Format



Data Payload (up to 24 phits)



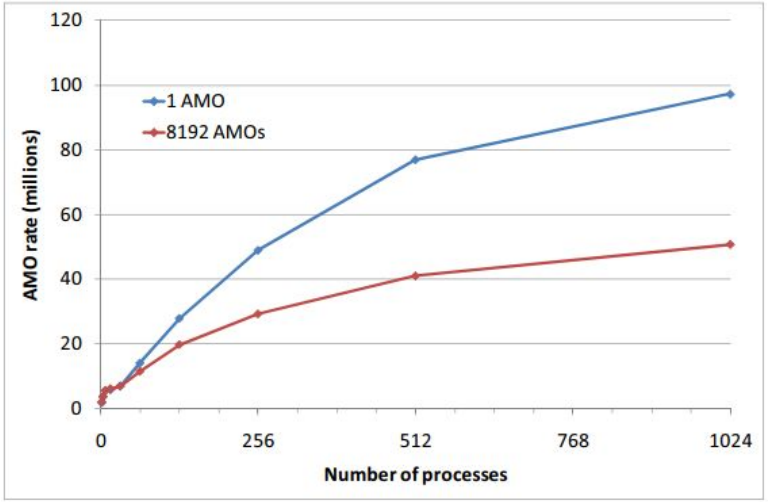
Fault Tolerance

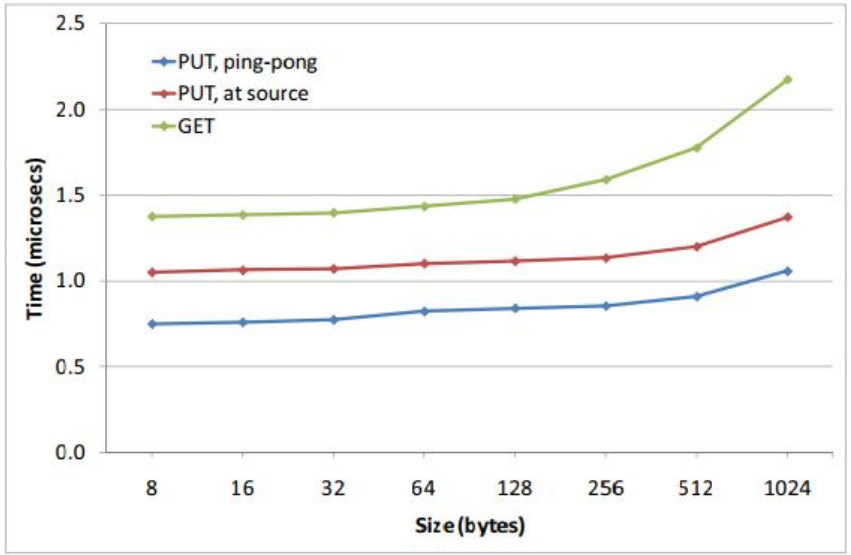
- 16-bit packet CRC
- 64 bytes of data + associated headers (96 bytes max)
- each torus connection comprises 4 groups of 3 lanes
- automatic link-level retry on error
- failure of a single lane
- complete failure of a link
- ECC for major memories and data paths protection
- completion events

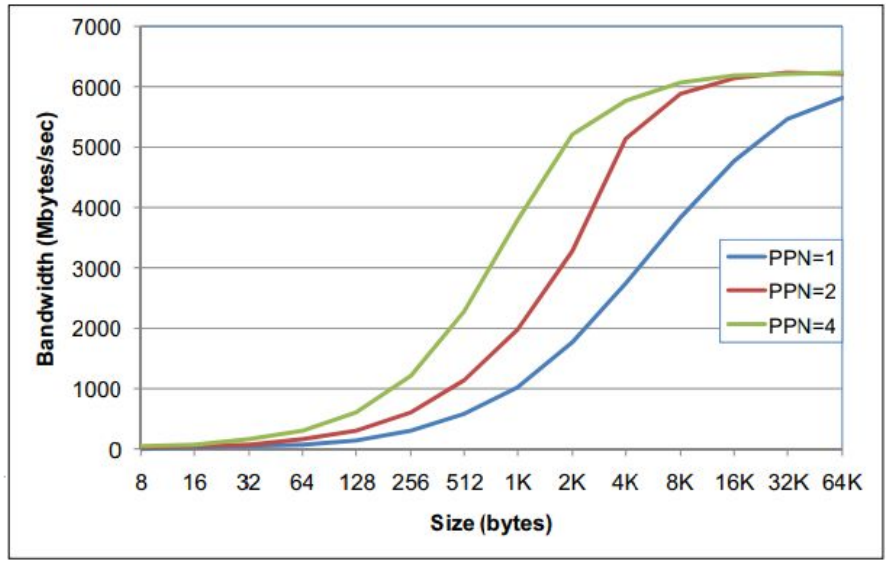

TITAN

Latency & Bandwidth

- On a quiet network
  - <1.5 us for small message (<=128 bytes)
  - Typically ~105 ns latency per hop







References

- https://www.cray.com/products/computing/xe-series
- https://www.cray.com/products/computing/xk-series
- https://de.wikipedia.org/wiki/Cray_XK7#/media/File:Titan_supercomputer_at_the_Oak_Ridge_National_Laboratory
- https://www.olcf.ornl.gov/wp-content/uploads/2013/02/Titan_Architecture_1-JL.pdf
- Abts, D. (2010). The Cray XT4 and Seastar 3-D Torus Interconnect.
- Alverson, B., Froese, E., Kaplan, L. and Roweth, D. (2012). Cray ® XC™ Series Net work.
- Alverson, R., Roweth, D. and Kaplan, L. (2010). The Gemini System Interconnect. *2010 18th IEEE Symposium on High Performance Interconnects*.
- Benner, A. (2012). Optical Interconnect Opportunities in Supercomputers and High End Computing.
- Inc, C. (2019). The Gemini Network.
- Larkin, J. (2012). *Titan Architecture*.
- Sun, Y., Zheng, G., Kalé, L., Jones, T. and Olson, R. (2012). A uGNI-based Asynchronous Message-driven Runtime System for Cray Supercomputers with Gemini Interconnect. *2012 IEEE 26th International Parallel and Distributed Processing Symposium*.
- Vaughan, C., Rajan, M., Barrett, R., Doerfler, D. and Pedretti, K. (2019). Investigating the Impact of the Cielo Cray XE6 Architecture on Scientific Application Codes.
- Vishnu, A., Bruggencate, M. and Olson, R. (2019). Evaluating The Potential of Cray Gemini Interconnect for PGAS Communication Runtime Systems.