



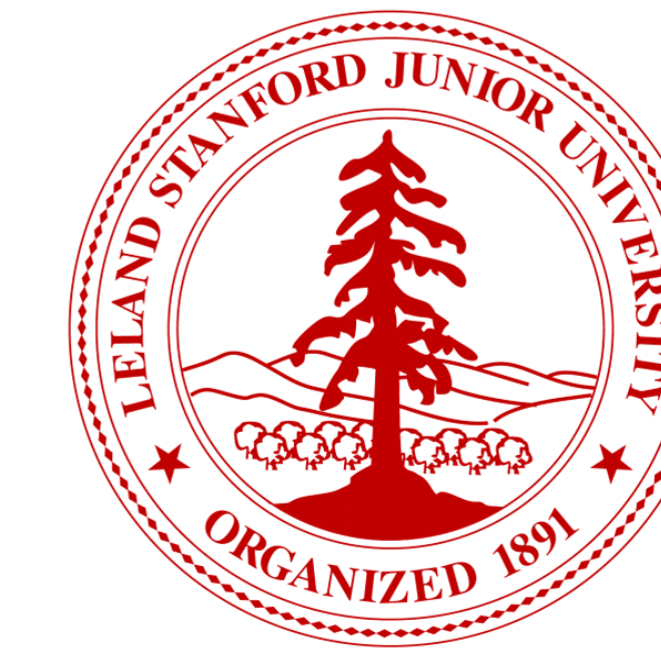
Merrimac Architecture

William J. Dally

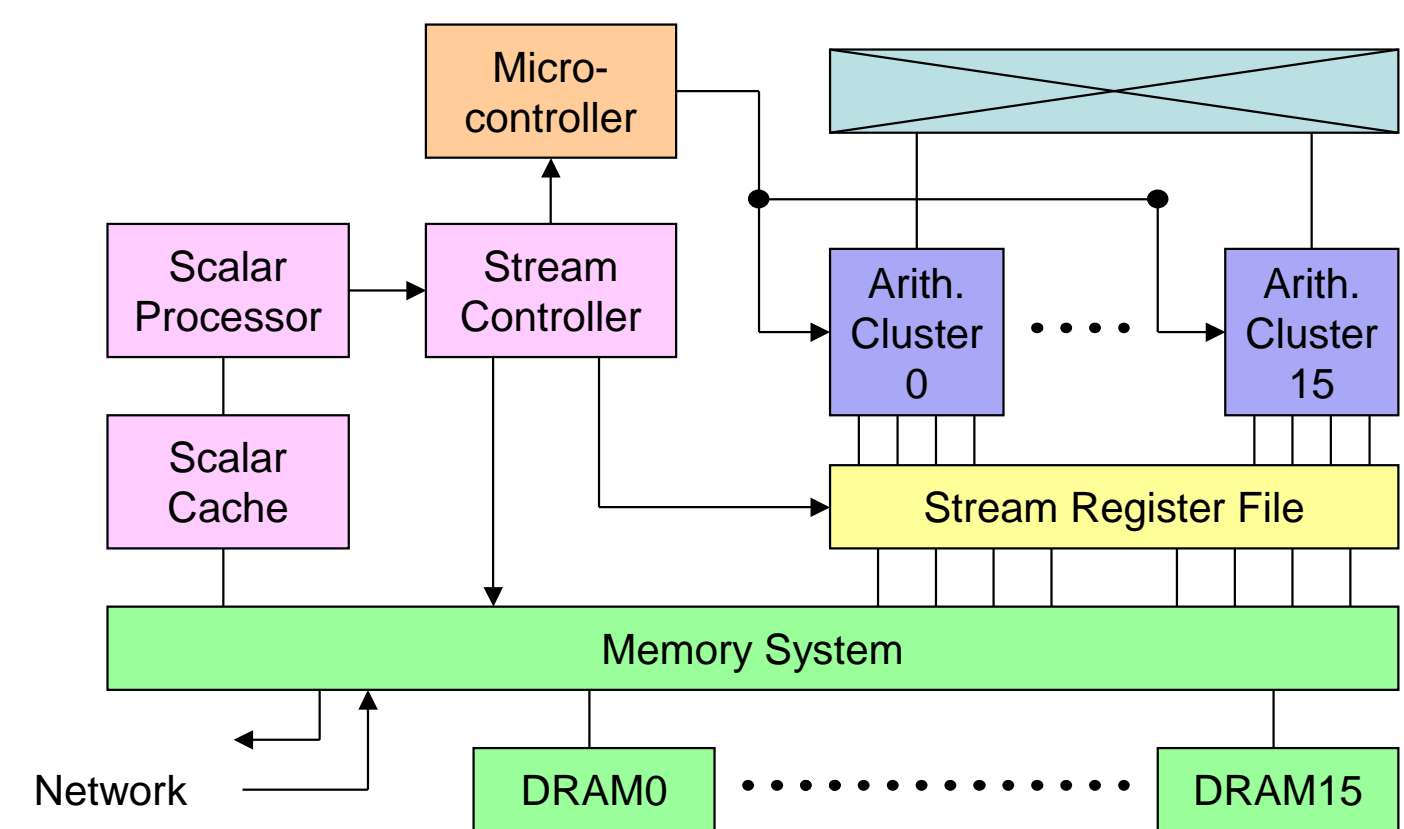
Mattan Erez
Nuwan Jayasena

Jung Ho Ahn
Ujval Kapasi
Abhishek Das

Francois Labonte
Timothy Knight

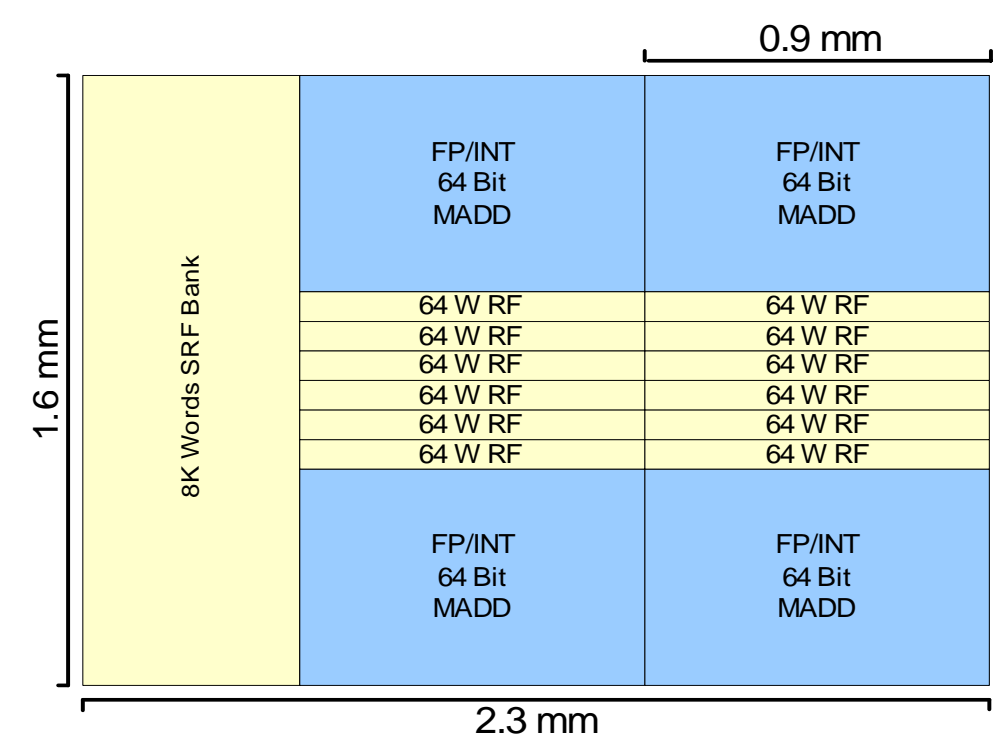


Merrimac Node



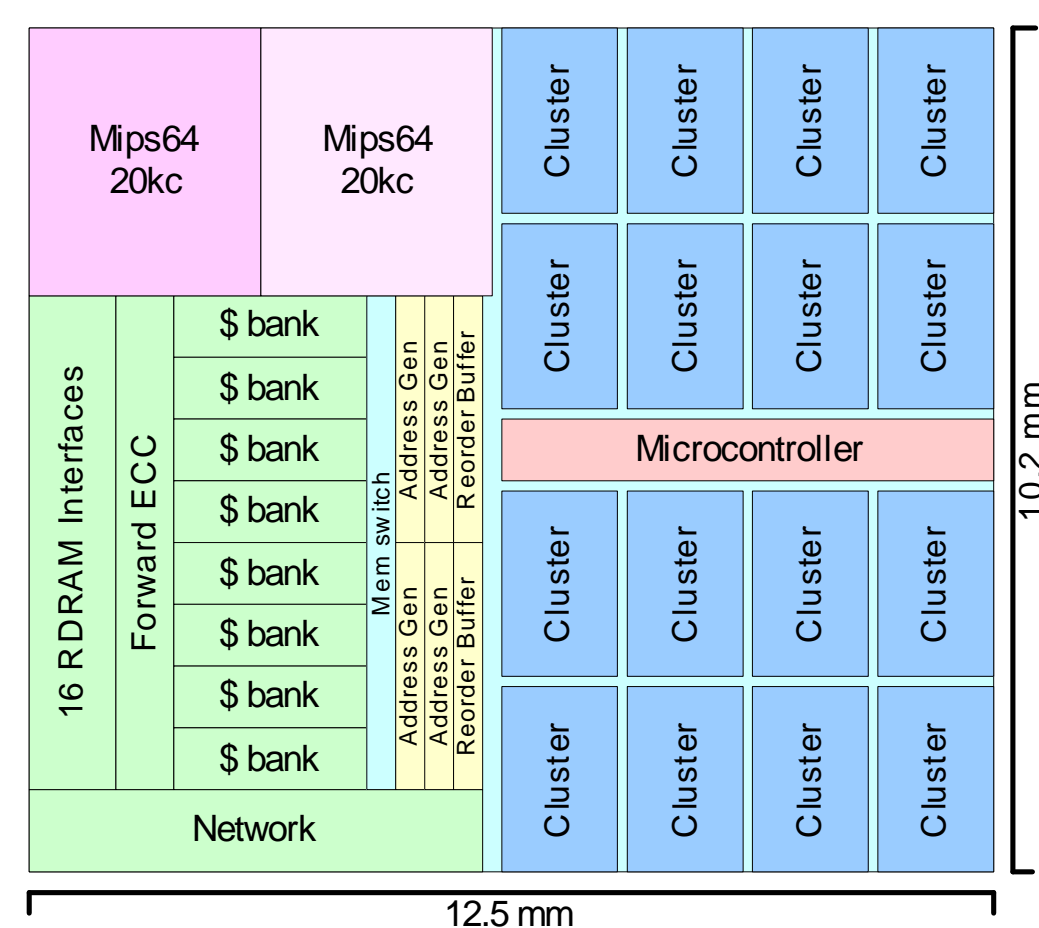
- 16 data-parallel compute clusters
- Integrated scalar processor
- Indexable SRF (poster)
- Capable memory system
 - 16 Gbyte/s (random access)
 - Stream cache
 - ScatterOp (poster)

Floorplan



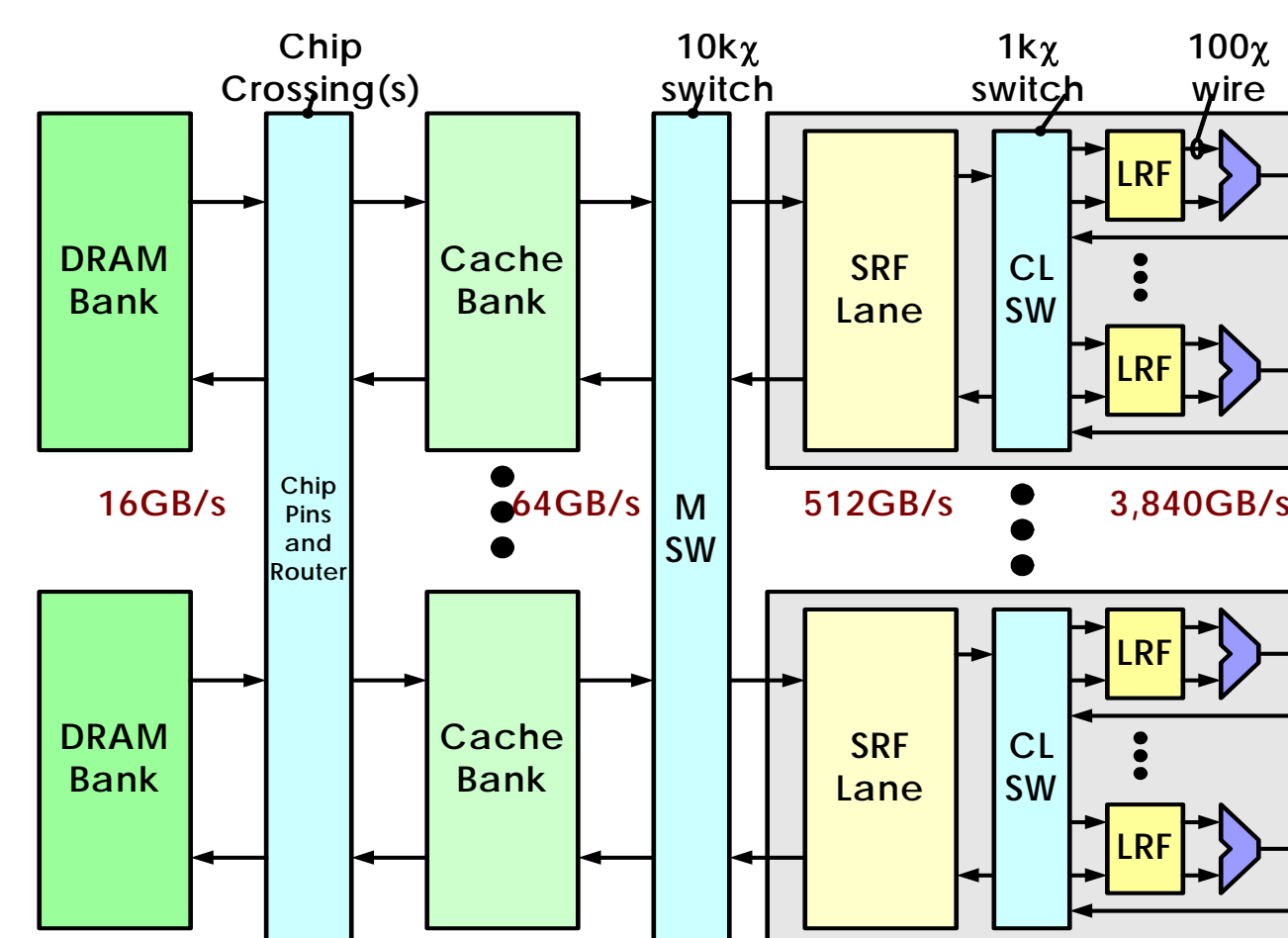
- 90nm tech (1 V)
- ASIC technology
- 1 GHz (37 FO4)
- 128 GOPs

- Inter-cluster switch between clusters
- 127.5 mm² (small ~12x10)
 - Stanford Imagine is 16mm x 16mm
 - MIT Raw is 18mm x 18mm
- 32 Watts (P4 = 75 W)



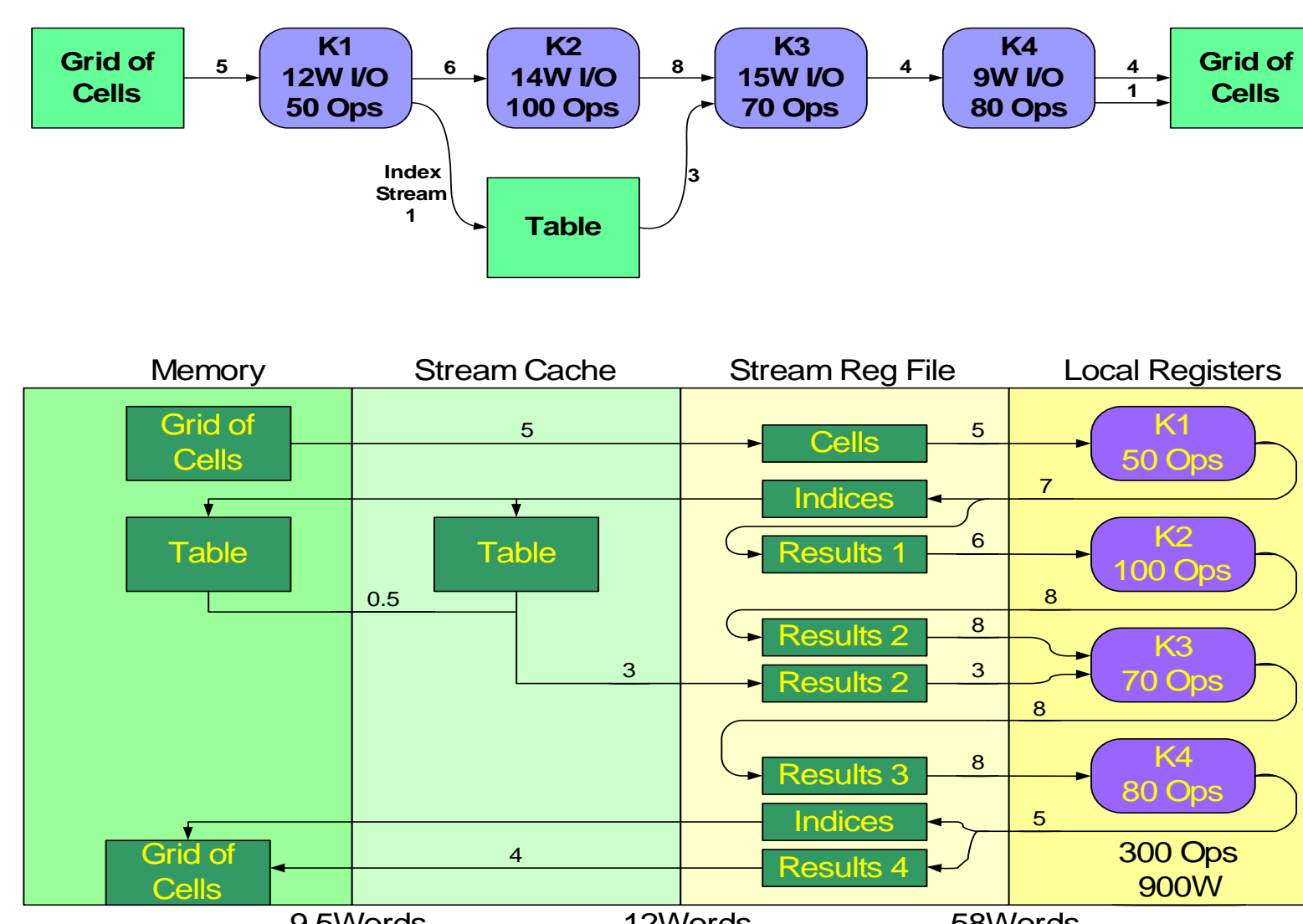
The Merrimac streaming supercomputer project aims to develop a scientific computer that offers an order of magnitude or more improvement in performance per unit cost compared to cluster-based scientific computers built from the same underlying semiconductor and packaging technology. We expect this efficiency to arise from two innovations: stream architecture and advanced interconnection networks. Organizing the computation into streams and exploiting the resulting locality using a register hierarchy enables a stream architecture to reduce the memory bandwidth required by representative computations by an order of magnitude or more.

Bandwidth/Register Hierarchy



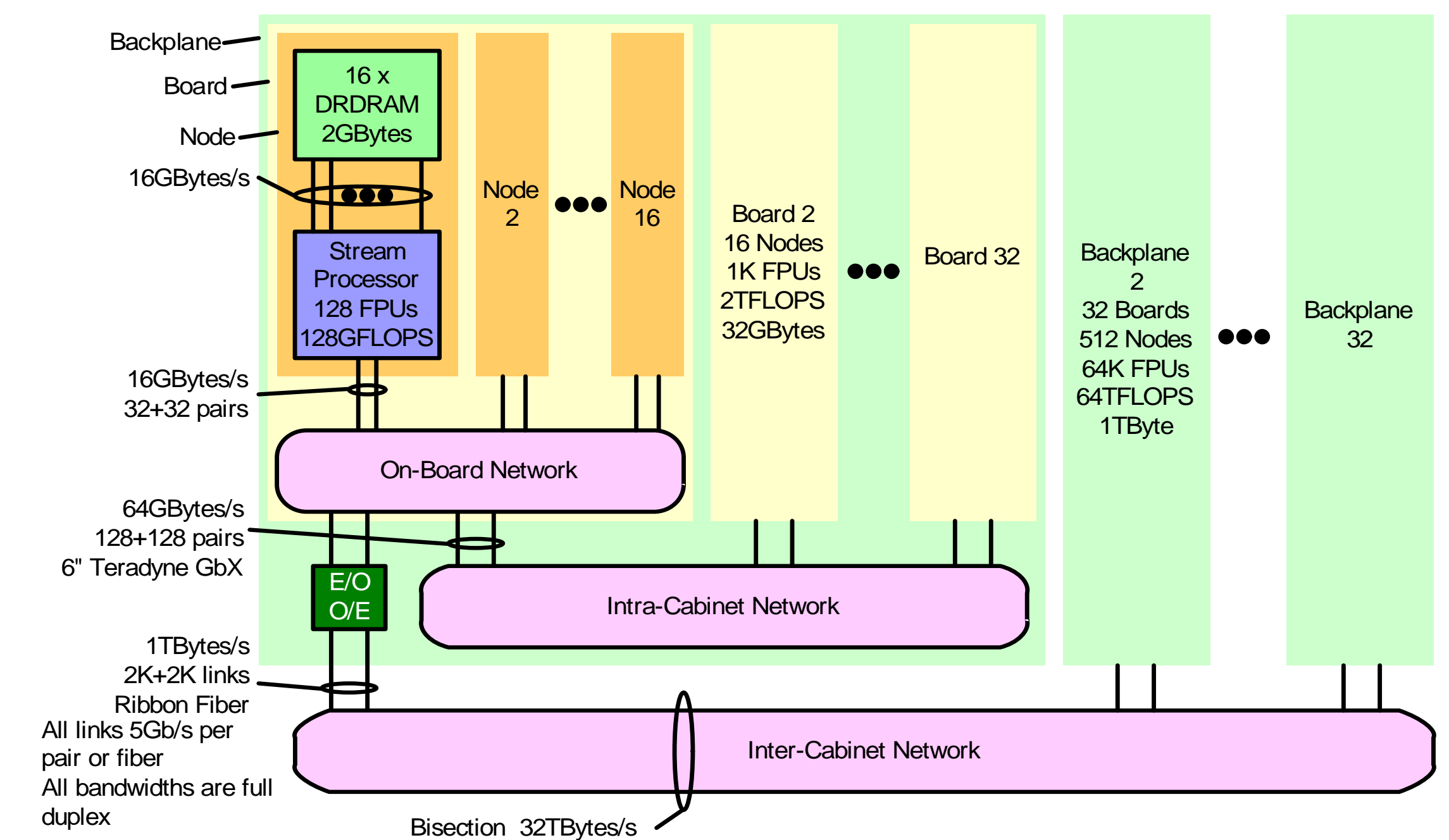
- Exploits locality
 - producer-consumer within kernel in the LRF
 - producer-consumer across kernels in the SRF
- Reduces the distance data travels
- Support large number of ALUs

Stream Applications

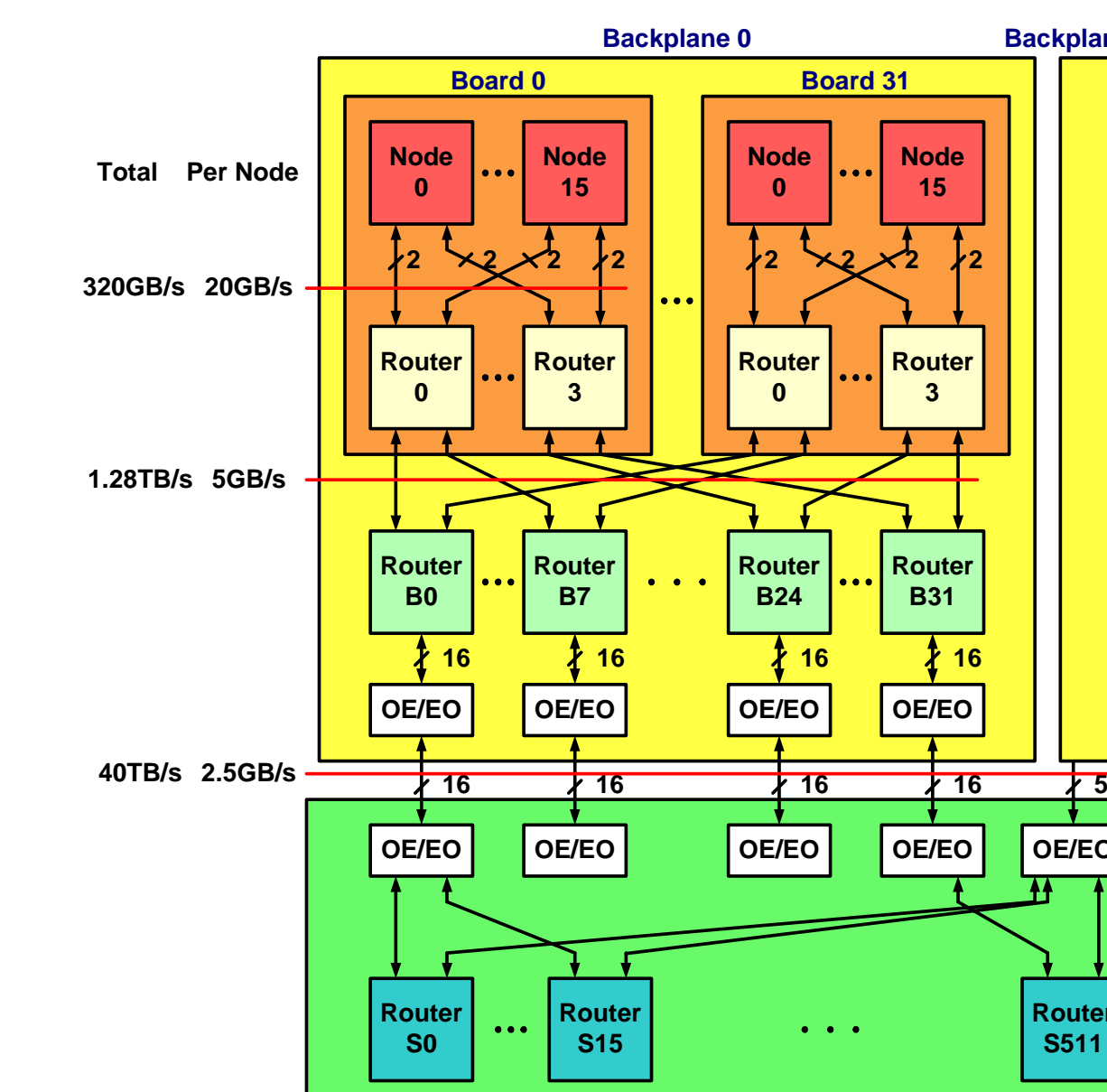


Application	Sustained GFLOPS	FP Ops / Mem Ref	LRF Refs	SRF Refs	Mem Refs
StreamFEM2D (Euler, quadratic)	32.2	23.5	169,505,648 (93.6%)	10,299,776 (5.7%)	1,354,448 (0.7%)
StreamFEM2D (MHD, cubic)	33.5	50.6	733,294,080 (94.0%)	43,762,752 (5.6%)	3,165,280 (0.4%)
StreamMD	23.3 will get -2X	14.3	427,743,216 (96.5%)	9,505,099 (2.1%)	5,978,848 (1.4%)
StreamFLO (kernel estimates)		50	(96%)	(2%)	(2%)

Merrimac System



Board and Network



- Flat memory bandwidth within a 16-node board
- 4:1 Concentration within a 32-node backplane, 8:1 across a 32 backplane system
- Routers with bandwidth B=640Gb/s route messages with length L=128b
 - Requires high radix to exploit

Reliability, Availability, and Serviceability

- Detect data errors using parity
 - Parity protect all arrays
 - DRAM, SRF, Cache, μcode store, Registers, Stream buffers, Reorder buffers
 - Parity protect buses
- Detect control and execution errors by replication
 - Replicate scalar core
 - Replicate micro-controller
- Use only half the clusters and address generators
 - Duplicate the computation
 - Compare the results
 - High-performance mode
 - Use all clusters, but less reliable

- Checkpoint - rollback
 - MTBF of 1e6 hours per board
 - Replace 1 board per month
 - MTBF of 1e8 per component
- Checkpoint duration ~5 minutes
- Recovery time ~10 minutes
- Checkpoint every 5 hours
- Slowdown of less than 3%

$$\text{slow down} = \frac{T_{cp,i} + T_{cp,d} + T_{cp,r}}{T_{cp,i}} \cdot \left(\frac{T_{cp,i}}{2} + T_r \right)$$

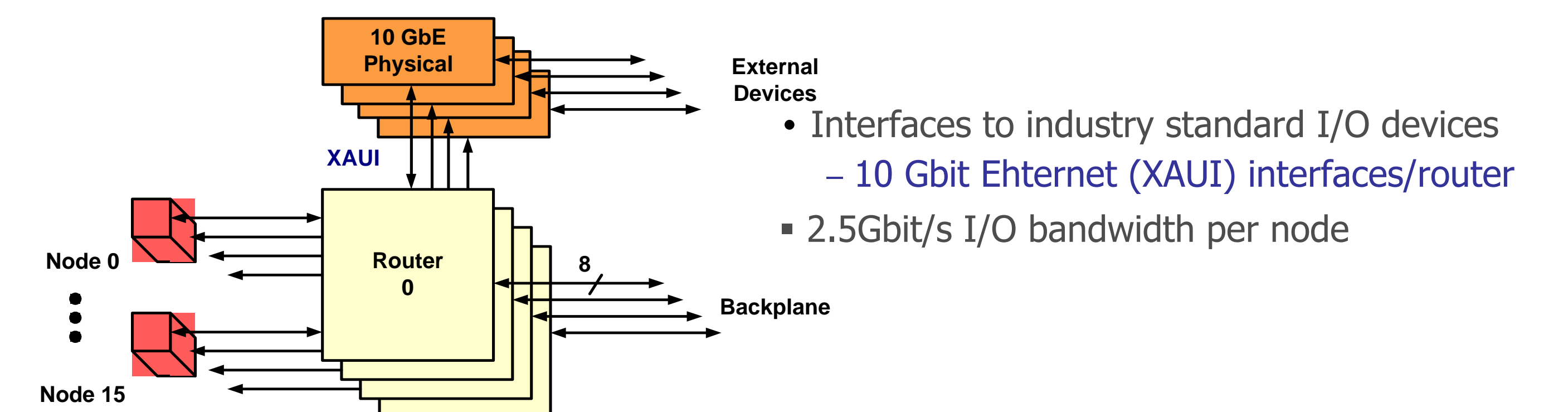
$$\frac{\partial \text{slow down}}{\partial T_{cp,i}} = -\frac{T_{cp,d} + T_{cp,r}}{T_{cp,i}^2} + \frac{T_{cp,i} + T_r}{T_{cp,i}}$$

$$\frac{\partial \text{slow down}}{\partial T_{cp,i}} = 0$$

$$\frac{1}{T_{cp,i}^3} + \frac{T_r}{T_{cp,i}^2} - T_{cp,d} = 0$$

- Redundant power and cooling
 - diode voting on boards
- Hot extra board per cabinet
- Network gracefully degrades

Input/Output



Future Research

- Interface between stream and scalar processor
 - short-stream effects
 - amount of software control
- Mechanisms for variable-rate streams
- Stream-cache studies
- Multi-node execution and simulation
 - stream scheduling
 - non-contiguous SRF