

Realizing High Arithmetic Intensity

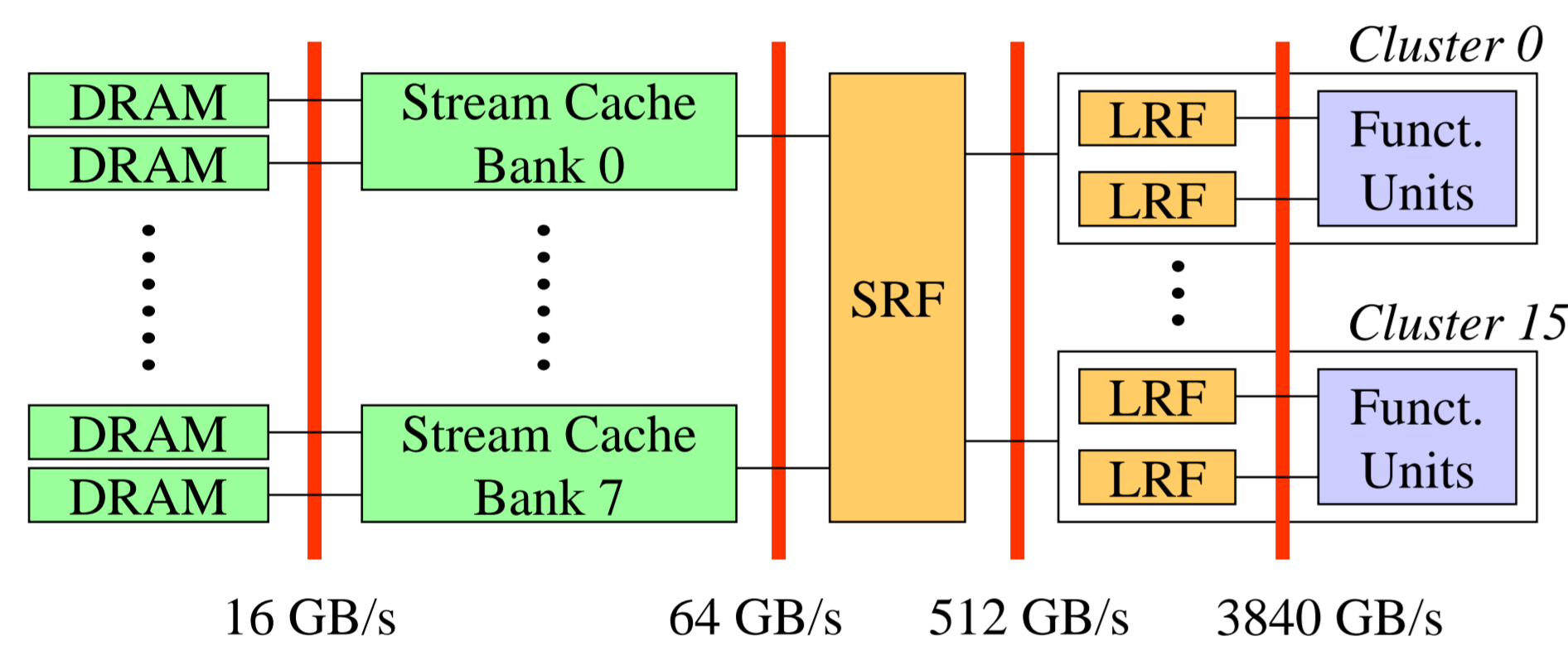
Stream Programs Specify Explicit Locality

- Locality of reference is exploited by providing a bandwidth hierarchy matched to application needs.
- Keeping data local reduces global bandwidth demands, enabling the ALUs to be kept busy rather than waiting on memory accesses.

Table 1: Application bandwidth demands

Application	Mem BW	SRF BW	LRF BW	GFLOPS
StreamFEM (Linear, MHD)	16 GB/s	72 GB/s	1491 GB/s	35
StreamMD				
StreamFLO				

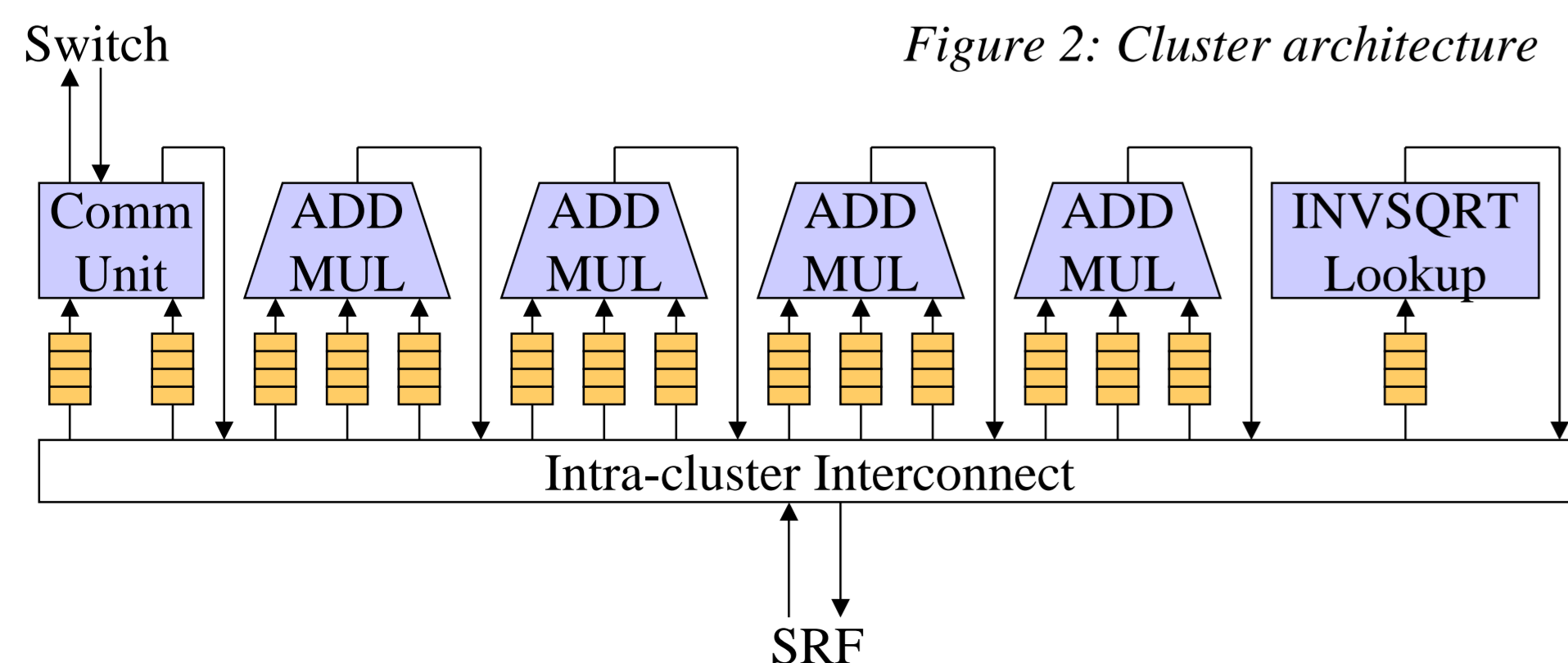
Figure 1: SSS bandwidth hierarchy



Stream Programs Specify Explicit Parallelism

- Data-level parallelism is exploited by using 16 dense ALU clusters which provide a total of 64 add-multiply 64-bit floating point units, yielding a peak performance of 128 GFLOPS.
- The clusters execute kernels in a SIMD fashion and are statically scheduled.

Figure 2: Cluster architecture



Streaming Supercomputer Node Architecture

Bill Dally (PI)
Timothy Knight
Jung Ho Ahn
Abhishek Das

Mattan Erez
Ujval Kapasi
Ben Serebrin
Nuwan Jayasena

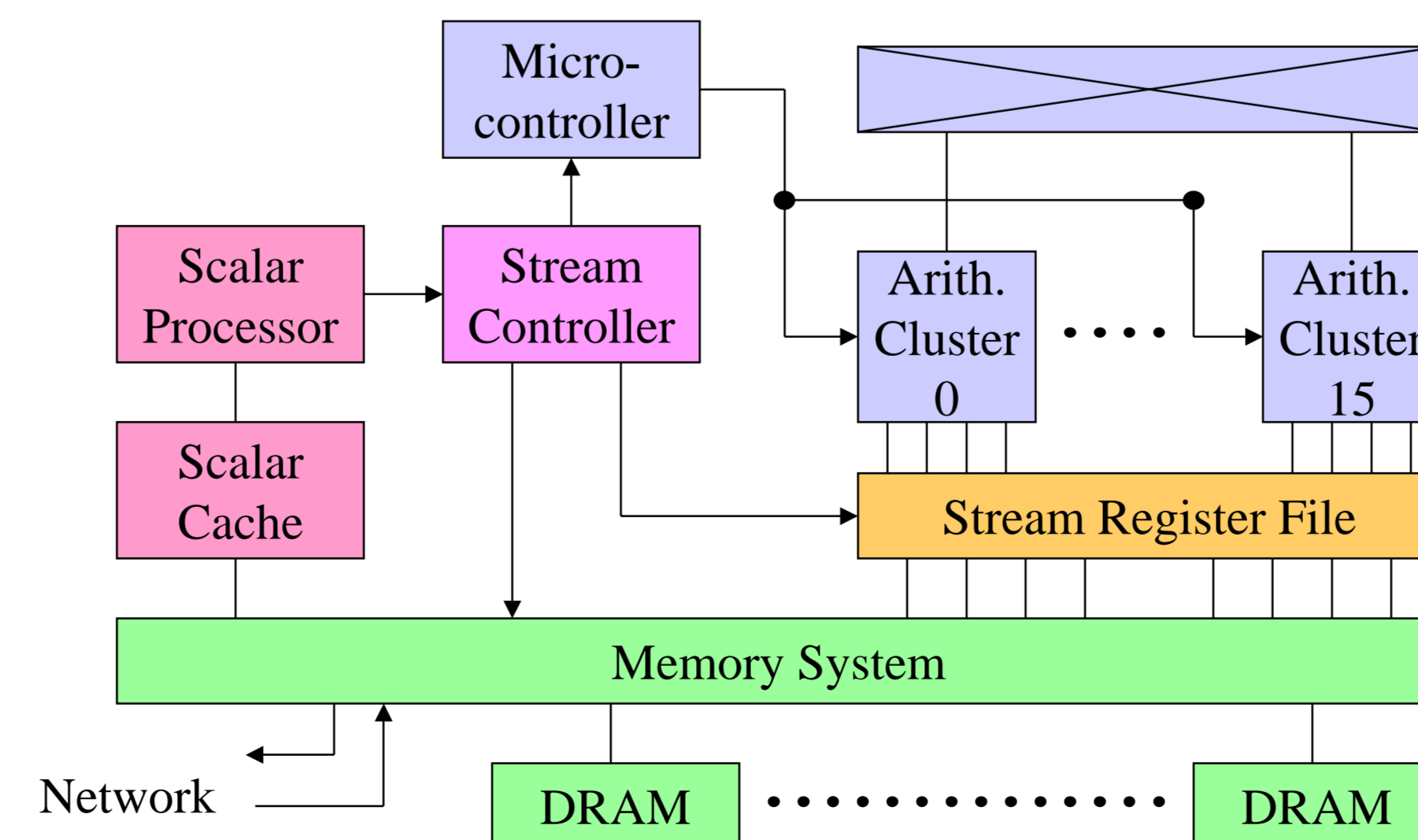
Achieves a 100x improvement in performance / cost over cluster- and vector-based supercomputer architectures by exploiting the stream model of computation.

To be done – justify the 100x statement.

I'm going to be looking at ASCII White, the Earth Simulator (NEC), and an extrapolated 2005 Pentium-based system to see if I can make a reasonable comparison.

Top-level Node Architecture

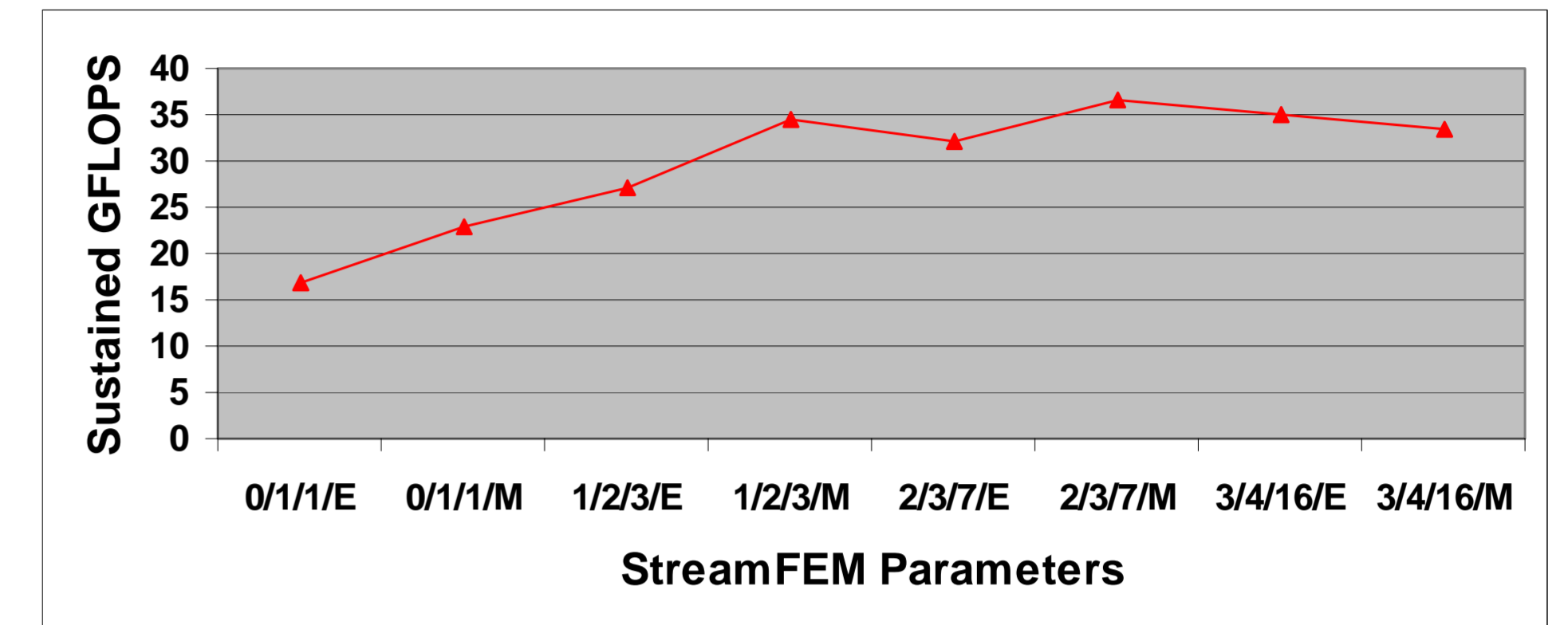
Figure 4: Streaming Supercomputer node architecture



Initial Results

- StreamFEM achieves a sustained performance of 33 to 36 GFLOPS * for quadratic or higher models.
* The simulator used for these results doesn't yet include the 3-operand add-multiple functional units, resulting in a peak of only 64 GFLOPS.

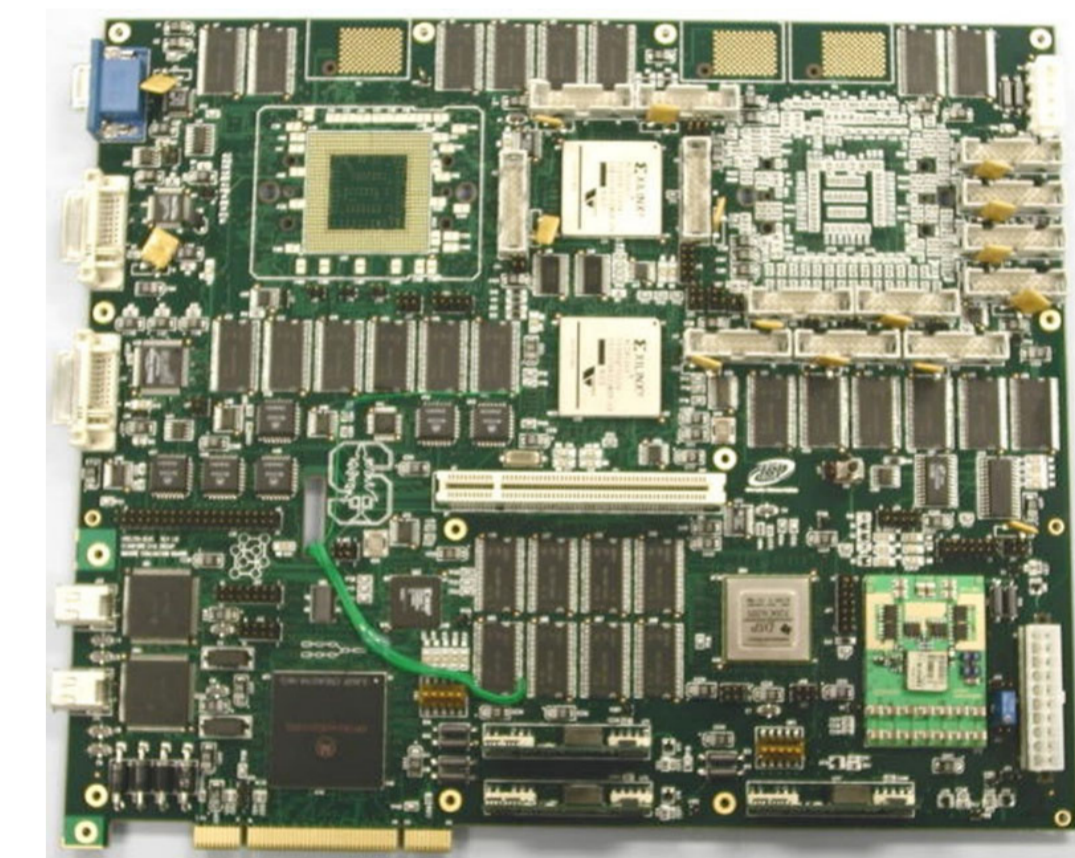
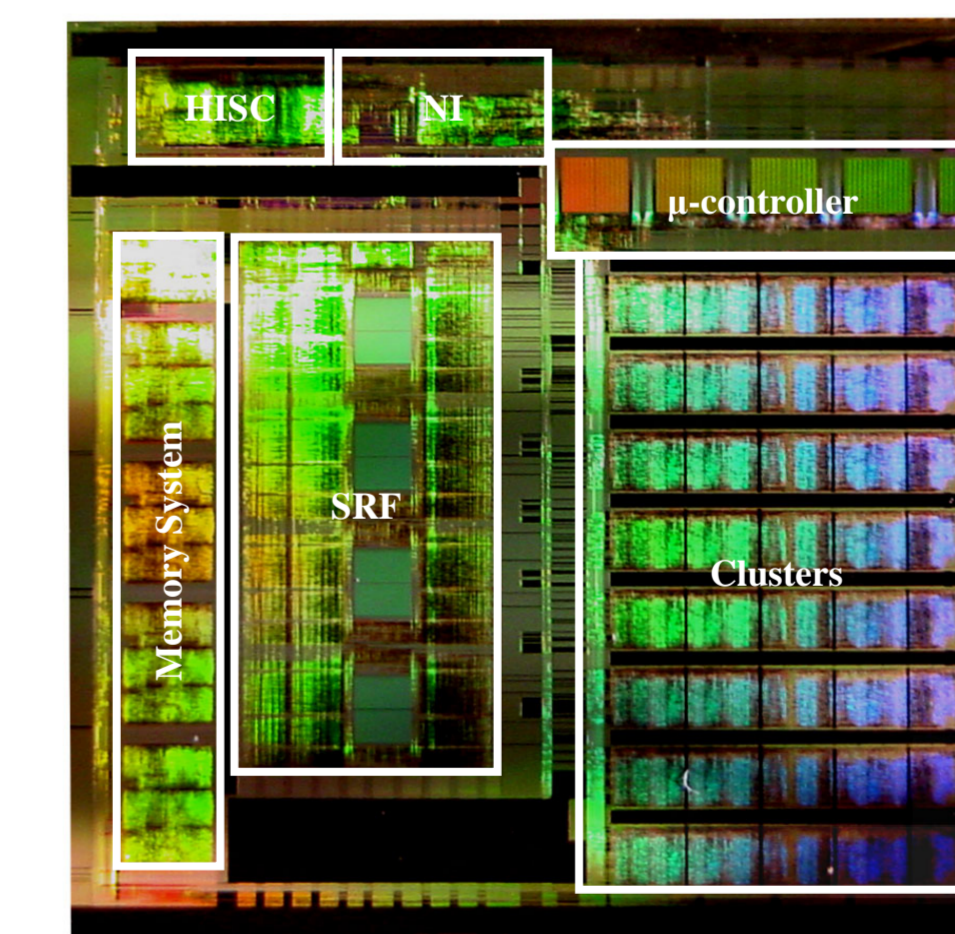
Figure 3: StreamFEM performance measurements



- Kernel schedules for StreamMD and StreamFLO indicate high ALU occupancy **** to be done – fill in actual kernel schedule data (Mattan?)**.

Imagine

- Prototype chip demonstrates the feasibility of the stream architecture.
- Samples are currently being run at 250 MHz for a peak performance of 10 GFLOPS.



Current Research Areas

- SRF indexing.
- Stream cache and memory system architecture.
- Scalar / stream integration.
- Cluster architecture (aspect ratio).