

Building More Powerful Less Expensive Supercomputers Using Processing-In-Memory

SAND2007-5840P



LABORATORY DIRECTED RESEARCH & DEVELOPMENT

Sandia National Laboratories

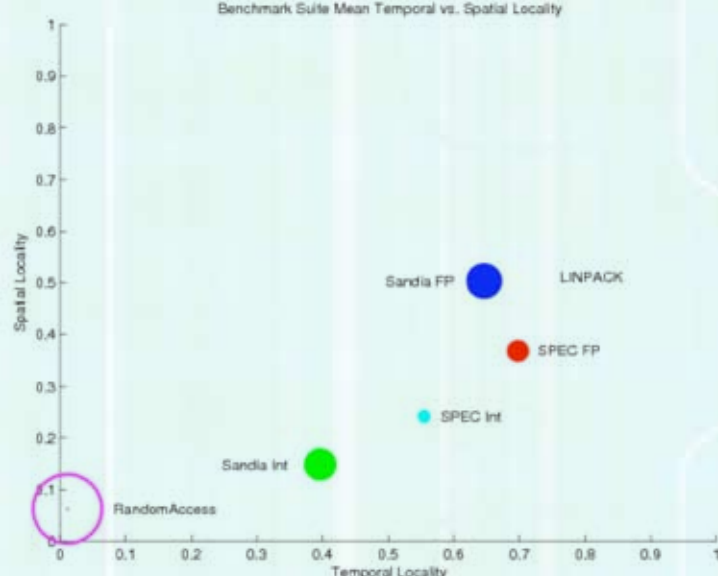
PI: Richard Murphy

Team Member: Arun Rodrigues



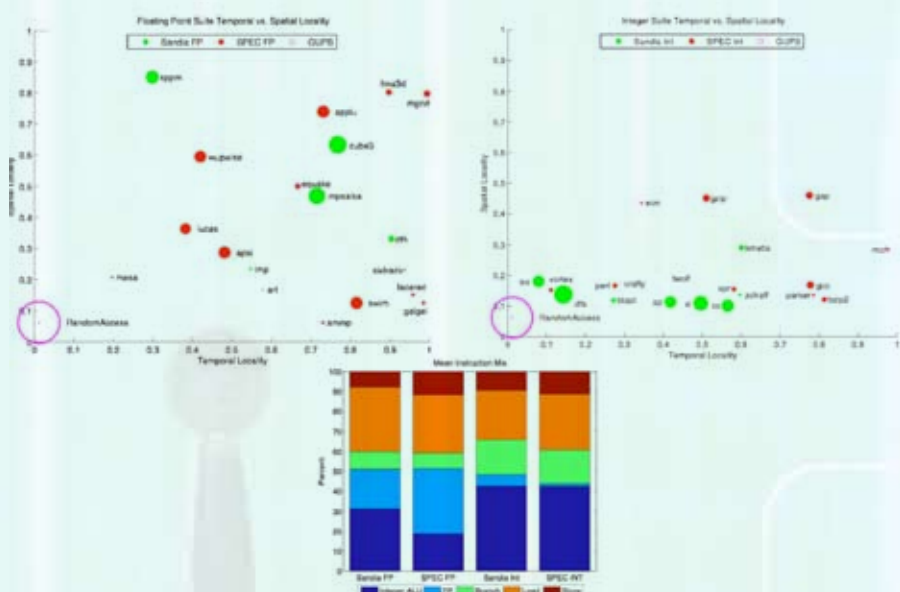
THE PROBLEM

Benchmark Suite Mean Temporal vs. Spatial Locality

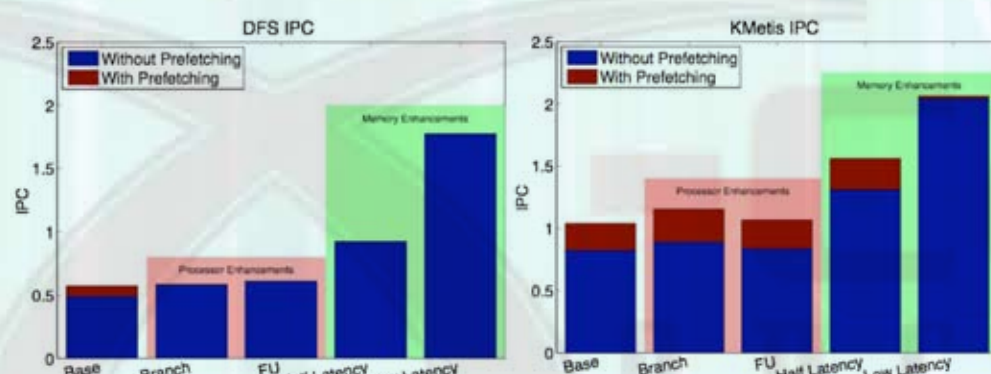


Applications (especially supercomputer applications) are dominated by memory performance

INDIVIDUAL APPLICATIONS COMPARED TO SPEC



FINDING THE BOTTLENECK: COMPUTATION, BRANCHES, OR MEMORY

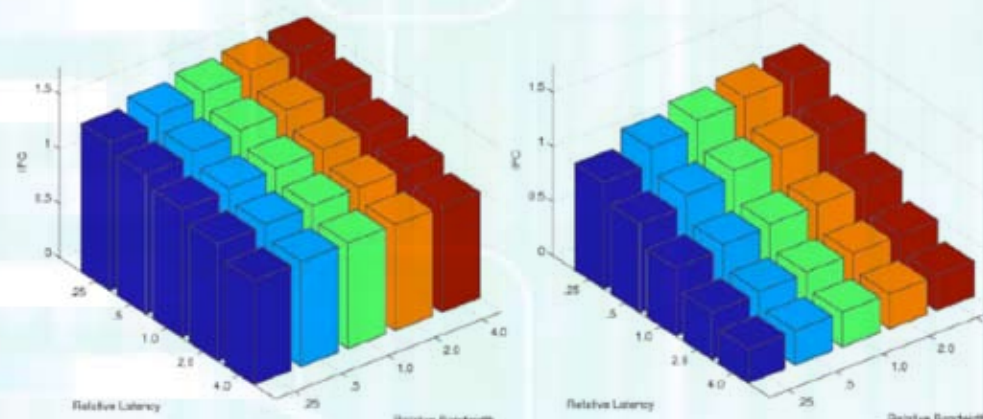


Even perfect branch prediction and infinite functional units would be less valuable than improving memory latency!

LATENCY/BANDWIDTH SENSITIVITY

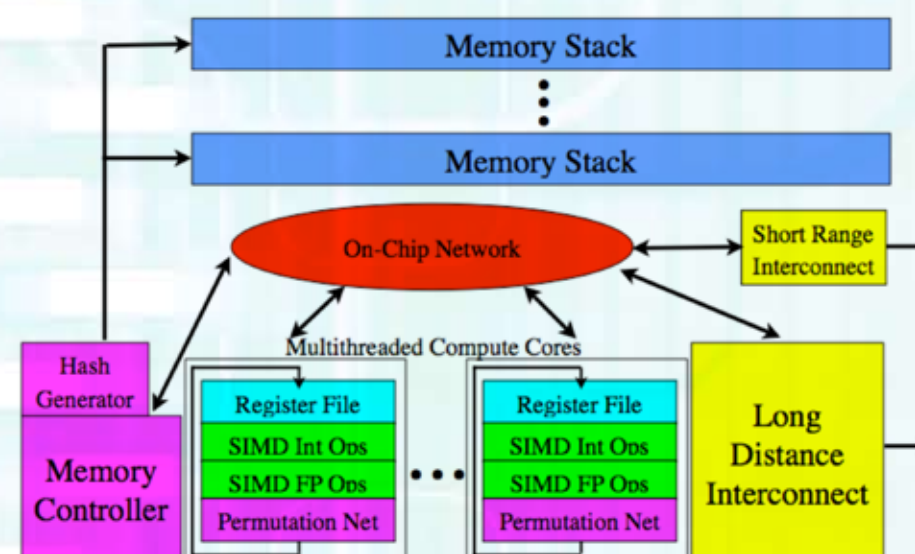
Floating Point

Integer



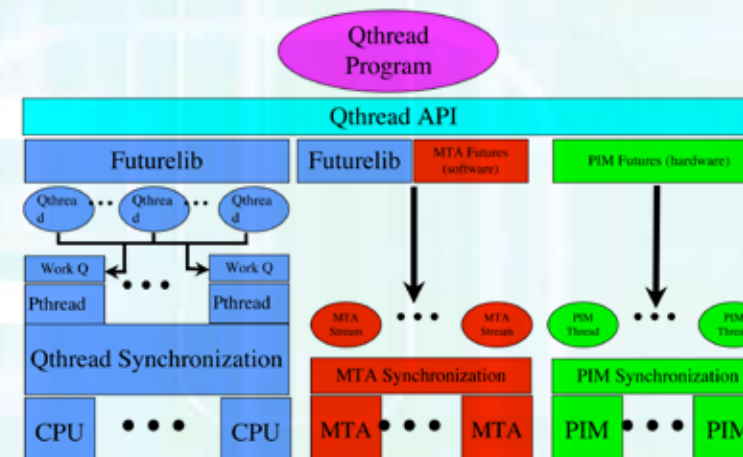
Emerging applications nearly 3x more sensitive to Latency and Bandwidth than traditional codes!

APPROACH: THE X-CALIBER ARCHITECTURE



Combines high-speed logic and dense DRAM in a 3D stack

EXPERIMENTAL RESULTS: DFS



- Up to 30% better than conventional with free coherency
- 2x-10x+ improvement over conventional!
- 4x+ better than an MTA running at 2 GHz

SIGNIFICANCE

- This project has the potential to create a new generation of supercomputers that are higher performance, lower power, and less expensive
- An outgrowth of this effort is joint work with a major memory vendor to define a new commodity memory part that is higher bandwidth, allows more outstanding memory references, and enables inexpensive 3D integration
- This work has quantified the impact of memory performance on a set of real Sandia applications and introduced a methodology for doing so with any other set of applications
- PIM simulation and Qthread runtime library developed
- 14 Applications Studied, 1 Journal Paper, and 3 Conference Papers in FY07



Sandia National Laboratories