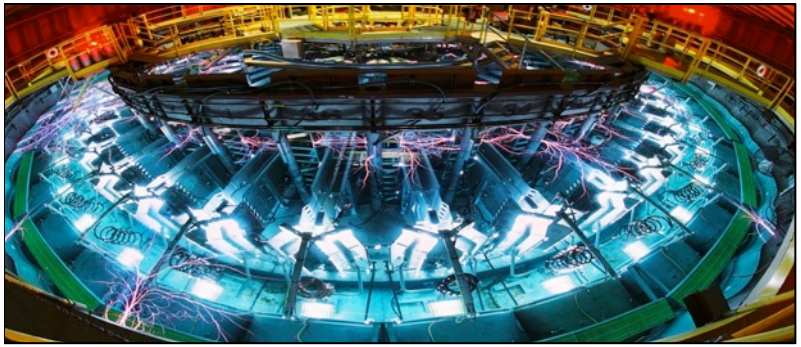
SAND2015-3462 C

Exceptional service in the national interest

Sandia National Laboratories



Scale-Free Memory Access Patterns

S.D. Hammond and Gwen Voskulien
Center for Computing Research, Scalable Computer Architectures
Sandia National Laboratories, NM
sdhammo@sandia.gov

U.S. DEPARTMENT OF ENERGY NNSA National Nuclear Security Administration

Sandia National Laboratories is a multi-program laboratory managed and operated by Sandia Corporation, a wholly owned subsidiary of Lockheed Martin Corporation, for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-AC04-94AL85000.

Acknowledgement

- Work is part of on-going collaboration with Trevor Mudge at Michigan, Ali Saidi and Andreas Hansson at ARM

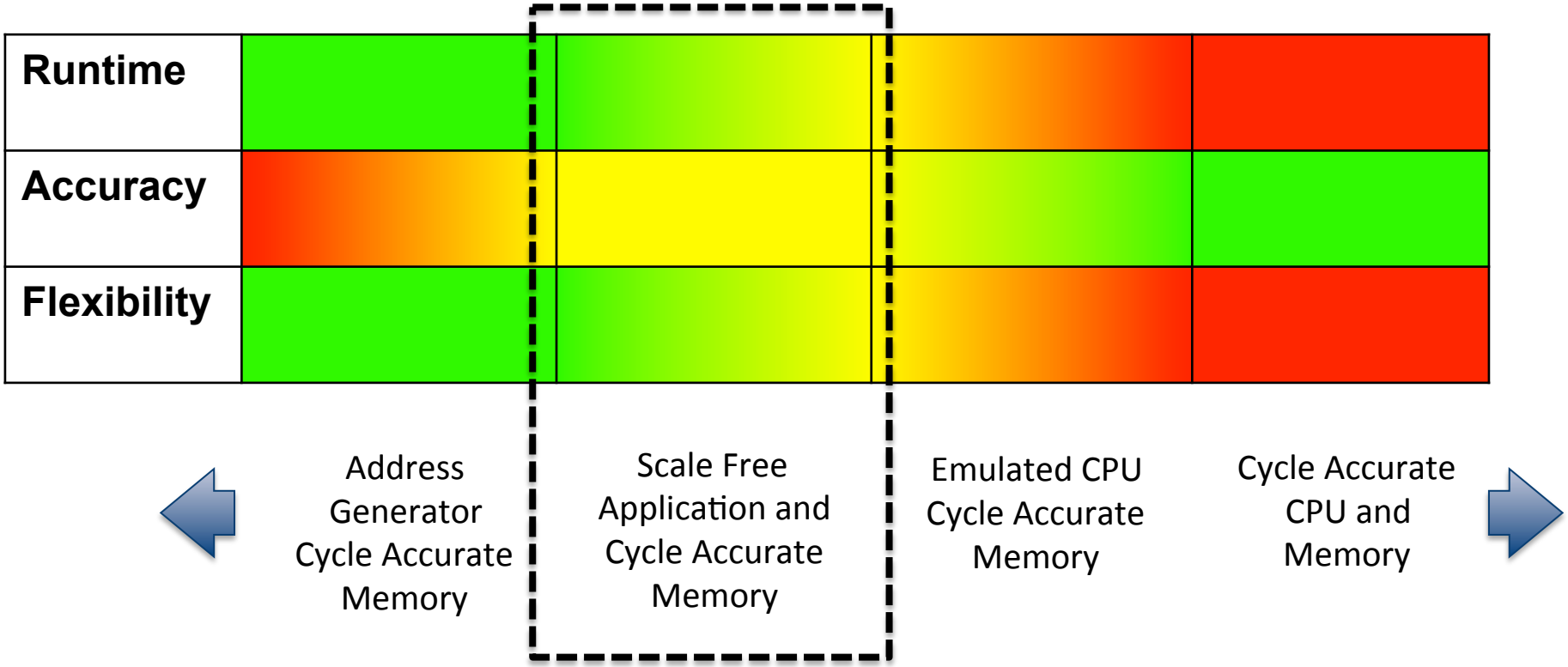
- Also part of on-going projects at Sandia to produce true scalable, flexible workload simulations

- Utilizes the Structural Simulation Toolkit (SST)
  - http://www.sst-simulator.org

It's all about the Memory…

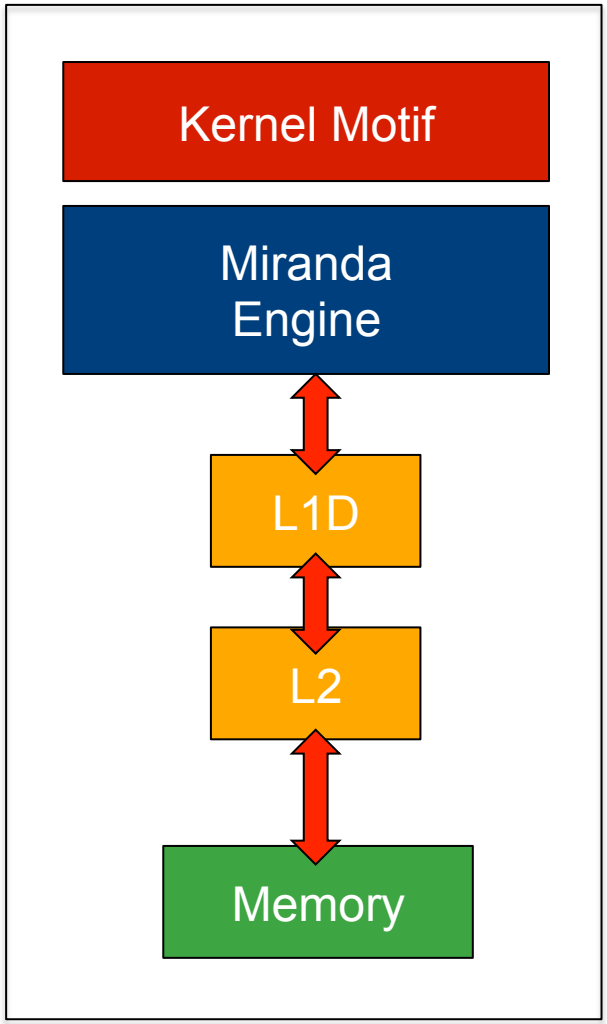
- Industry delivered incredible improvements in floating point and mathematical operation support
- But … has not delivered memory that can sustain supplying operands
- Pushing us towards future complex memory systems and technologies
  - HBM, HMC, Non-volatile and DRAM, possibly off-node burst buffers
- Where *much* of the strain will be in getting applications to perform
  - Even moderately threaded and vectorized code can make more demands than memory can supply
- Need a *much* better understanding of how memory performance will scale

Goldilocks of Application Models



What are we doing?

- Built a kernel "motif" as an address generation pattern
  - Encodes operations ***and*** dependencies
- Execute out-of-order/in-order operation issue in simulated memory hierarchy
- Full NoC, cache, prefetcher and memory models
- Analysis for traffic patterns, NoC pressure, prefetch strategy *etc* in parallel
- Fast and scalable (Large GBs, > Bn operations)

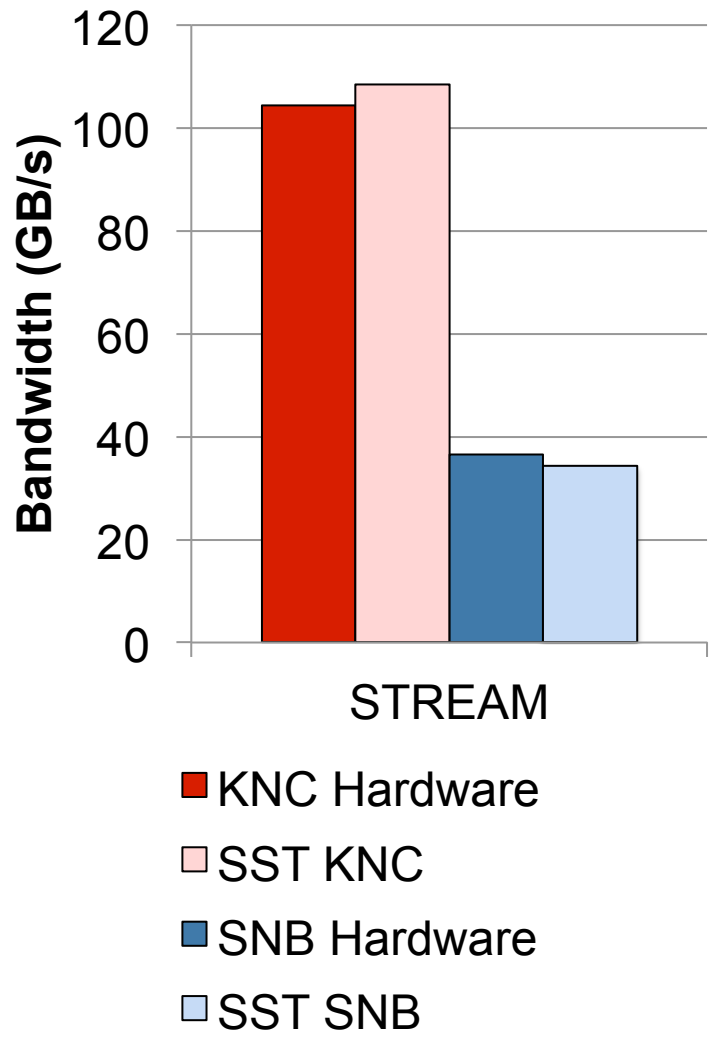


Kernel Workload at DOE

- Threaded Microbenchmarks
  - STREAM
  - GUPS
  - True random read/writes
- Threaded Application kernels
  - Vector dot-product
  - AXPY
  - Sparse Matrix Vector Multiplication
  - Dense 2D/3D stencil operations (variable stencils)
- Can arrange multiple kernels to look like larger applications
- Kernel motifs flexibly parameterized to represent wide variety of apps

Memory Models

- Using kernel motifs to drive large scale studies of:
  - Multiple levels of memory
  - Non-volatile memory performance
  - Prefetch strategies for different memory classes
  - .. Reliability schemes?
- Trying to inform design of primitives in mathematics libraries and up into program design



Sandia National Laboratories

Exceptional service in the national interest