Compute-Near-Memory Architecture

Glen Edwards and Tony Brewer
Micron Technology, Inc.

New Materials and Devices: Framework for Novel Compute (FRANC)

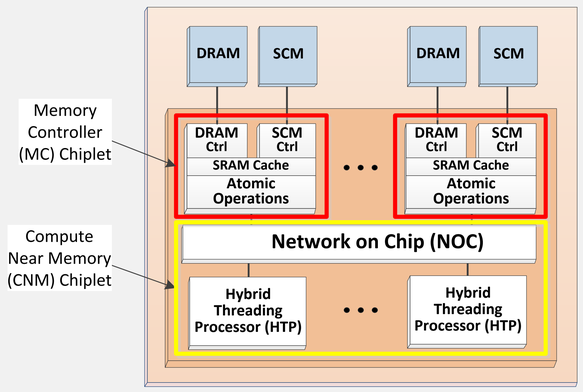
The Challenge

Sparse data sets that greatly exceed a processor cache size are a challenge for most systems

- Processors are typically optimized for high cache hit rate (>90%)
  - Low cache hit rate results in idle cores
- Memory accesses are cache line size (64B)
  - Sparse data sets result in memory accesses where the majority of accessed data is not used

Compute-Near-Memory (CNM) Architecture

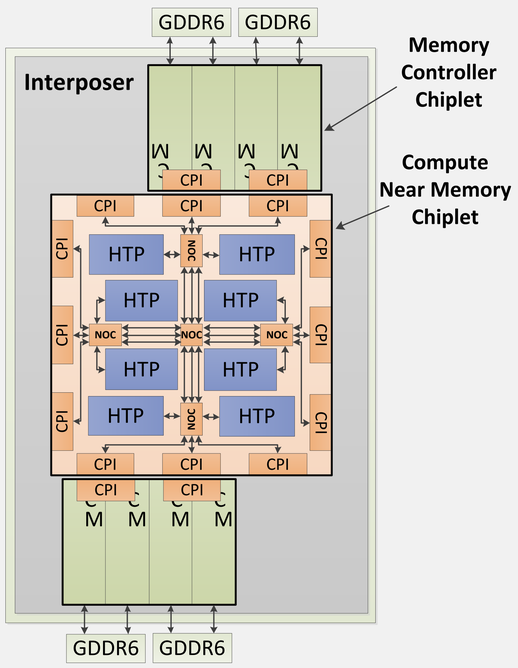
- Integrates compute elements with high-performance memory
- Latency-tolerant Hybrid Threading Processor (HTP) tightly integrated with memory system
- Chiplet-based implementation for flexible configurations
- Cycle-based simulation model for power and performance estimation of applications



HYBRID THREADING PROCESSOR (HTP)

- RISC-V ISA (RV64G) with extensions for thread and message management
  - Thread Management (Thread Create, Return, Join)
  - Message Management (Message Send, Broadcast, Receive, Listen)
  - Non-Cached Loads and Stores (Integer and Float)
- High thread count barrel processor
  - Similar to Cray's MTA architecture
  - One instruction per thread per scheduling interval (avoids register hazard checking)
- Event driven processor
  - Pause for memory response
  - Pause for thread join
  - Pause for message reception
- Efficient memory usage
  - Memory access size 8, 16, 32 or 64B
- Software managed coherency
  - Small cache per thread
  - Atomics performed at memory

1x1 CNM Configuration



Graph Spectral Clustering Application

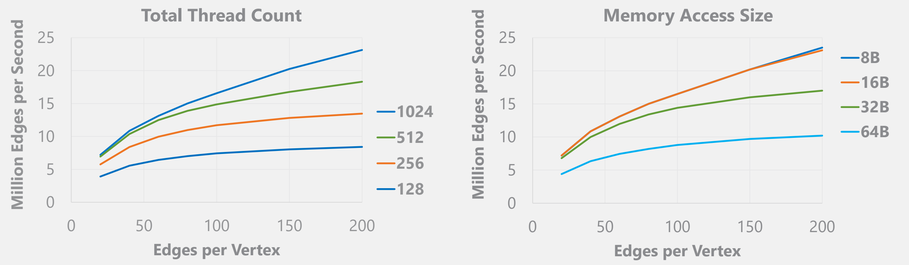
GRAPH COMMUNITY DETECTION USING SPECTRAL METHODS

- Uses linear algebra to compute eigenvalues for the adjacency matrix associated with a graph
- Lowest eigenvalues can be used to partition the graph
- Sparse data structures store the graph vertices, edges and properties

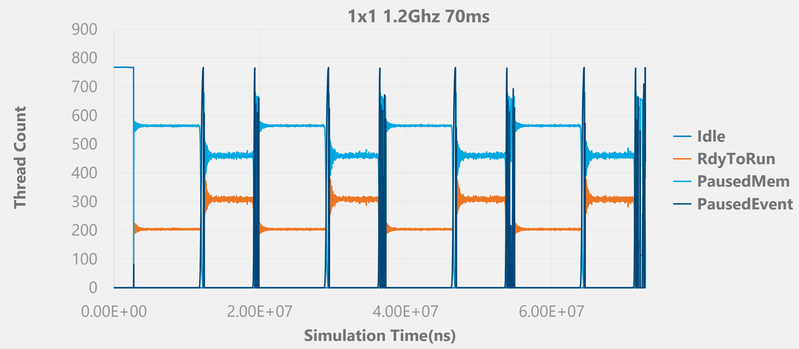
- Profile on an X86 system

Overhead	Symbol
13.82%	[.] svd_ATxb
13.36%	[.] svd_ATxb2
10.70%	[.] svd_Axb
10.01%	[.] substruct
9.93%	[.] svd_Axb2
6.59%	[.] _IO_vfscanf

SENSITIVITY ANALYSIS TO DETERMINE OPTIMAL CONFIGURATION



THREAD STATE MONITORING PROVIDES INSIGHTS INTO RUN-TIME DYNAMICS

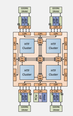


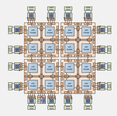
COMPARISON TO REFERENCE PLATFORMS

Haswell 8-threads 1-socket	Nvidia K80 (Host + 1 GPU)	Nvidia DGX-1 (Host + P100) 1 P100 time	NOC 1x1 Config 1.2Ghz Simulated	NOC 2x2 Config 1.2Ghz Simulated
13.5 Sec	5.0 Sec[1]	3.95 Sec[1]	2.90 Sec	0.814 Sec
140 Watts	340 Watts	(Note: system at a	23.8 Watt	90.9 Watt
1890 Joules	**1703 Joules**	cloud provider, no	**69 Joules**	**74 Joules**
27.4x	**24.7x**	power info)	**1.0x**	**1.07x**

Notes:

1. Times reported for GPUs do not include time to copy graph from host to GPU (120 sec).





This research was developed with funding from the Defense Advanced Research Projects Agency (DARPA). The views, opinions and/or findings expressed are those of the author and should not be interpreted as representing the official views or policies of the Department of Defense or the U.S. Government

THE ELECTRONICS RESURGENCE INITIATIVE