Graph Analysis and Novel Architectures

Jason Riedy (all opinions my own, no plans)
Lucata Corporation / Emu Technology
Sparse Days, 24 November 2020

Monument aux Combattants de la Haute-Garonne

Graph Analysis v. Hardware Architecture

“We” want:

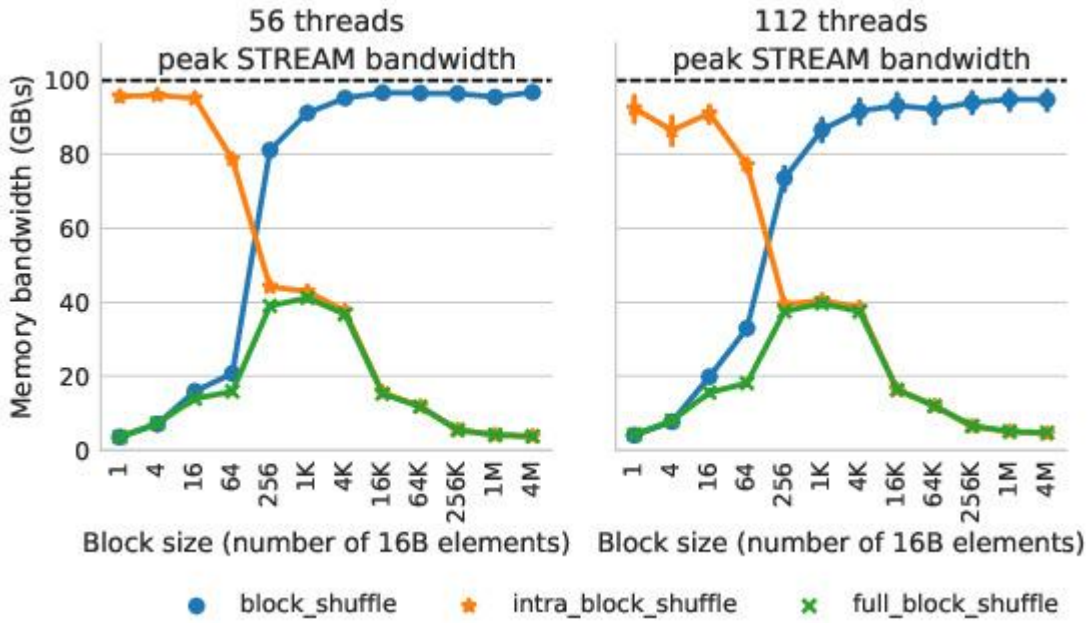
- Fine-grained memory access,
- fine-grained synchronization,
- sane floating-point (to be defined someday), and
- everything else that drives HW people nuts.

WHY NOT?

Graph Analysis v. Hardware Architecture

"It's too hard." Need wide memories, big cache lines, *etc*.

Nope.

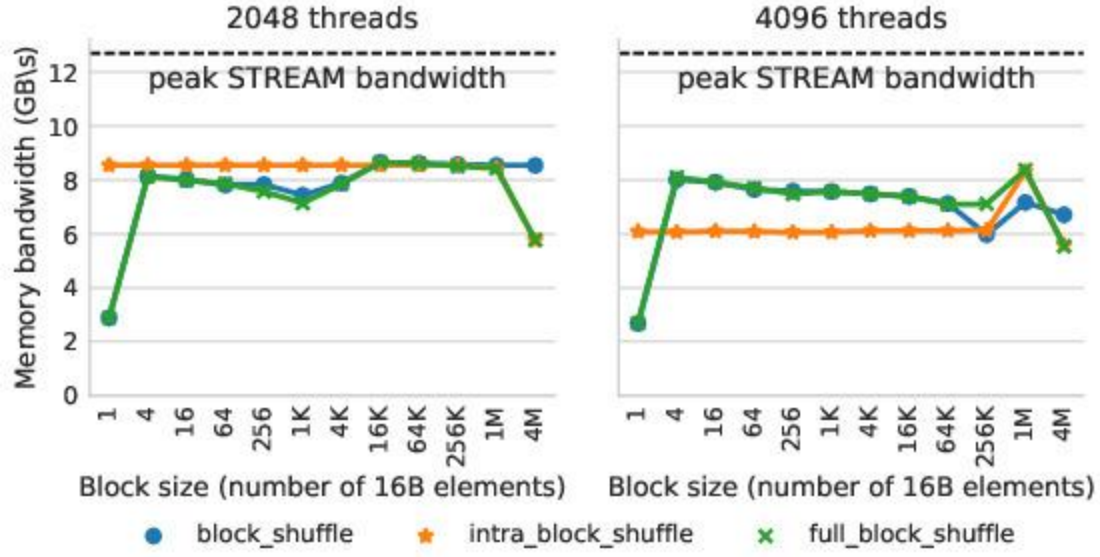


Jeffrey Young, Eric Hein, Srinivas Eswar, Patrick Lavin, Jiajia Li, Jason Riedy, Richard Vuduc, and Thomas M. Conte. A Microbenchmark Characterization of the Emu Chick. *Parallel Computing*, September 2019. DOI 10.1016/j.parco.2019.04.012.

Graph Analysis v. Hardware Architecture

"It's too hard." Need wide memories, big cache lines, *etc*.

Nope.

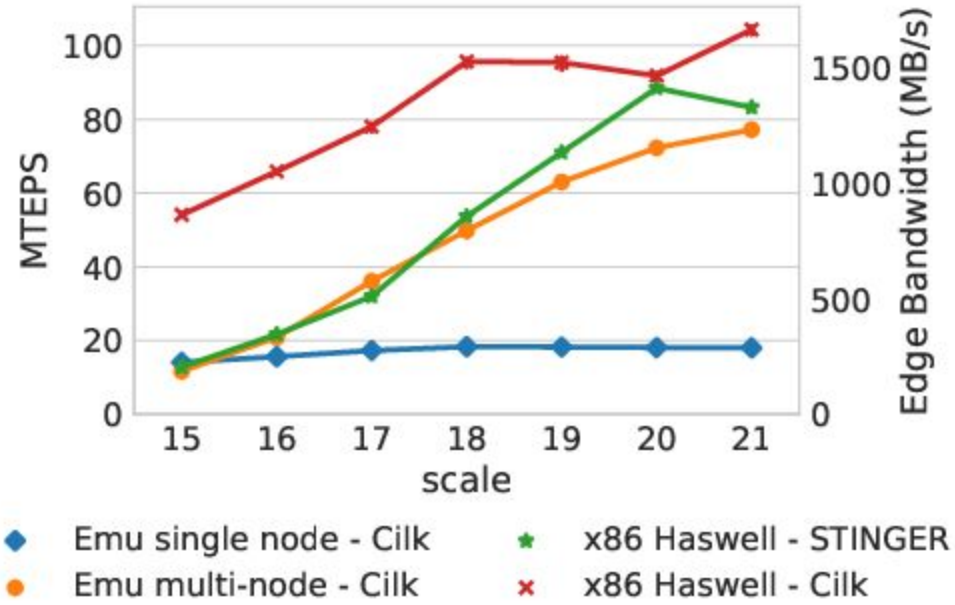


Jeffrey Young, Eric Hein, Srinivas Eswar, Patrick Lavin, Jiajia Li, Jason Riedy, Richard Vuduc, and Thomas M. Conte. A Microbenchmark Characterization of the Emu Chick. *Parallel Computing*, September 2019. DOI 10.1016/j.parco.2019.04.012.

Graph Analysis v. Hardware Architecture

"It's too hard." Need wide memories, big cache lines, *etc*.

Nope.

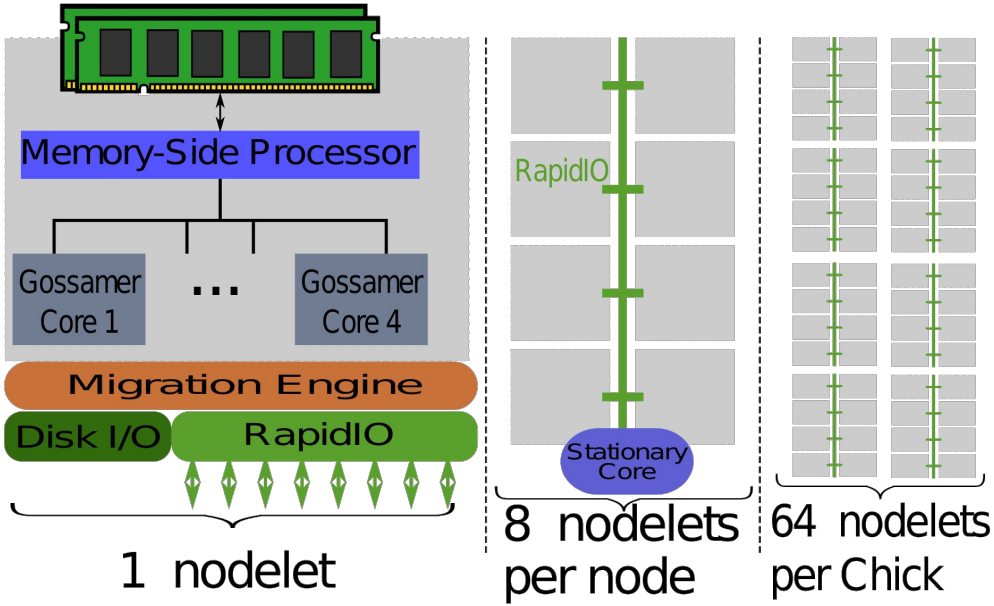


Jeffrey Young, Eric Hein, Srinivas Eswar, Patrick Lavin, Jiajia Li, Jason Riedy, Richard Vuduc, and Thomas M. Conte. A Microbenchmark Characterization of the Emu Chick. *Parallel Computing*, September 2019. DOI 10.1016/j.parco.2019.04.012.

How? Being specific.

The Lucata / Emu architecture focuses on fine-grained memory access.

This really exists. And is PGAS. Because...

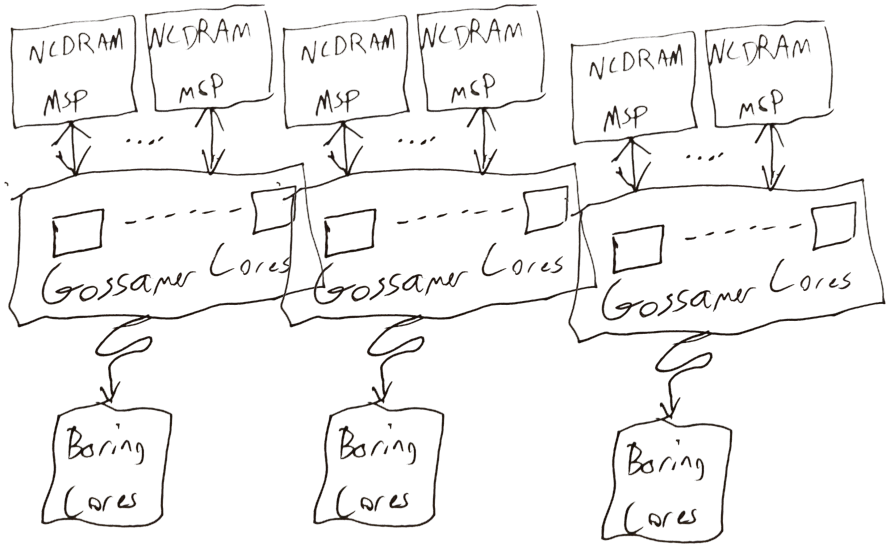


- No cache.
- The OS is handled by the “boring” part.
- Physically distributed memory.
- Many threads to tolerate…
- LOCAL LATENCIES.
  - Read remotely? MIGRATE.
  - Small context, one flit.
  - Plenty of references.
- Oh, and by the way…
  - Narrow channel DRAM: No wasting cache lines (so not using ⅛ BW).
  - Memory-side processing.
  - Including floating-point accumulation.

How? Being specific.

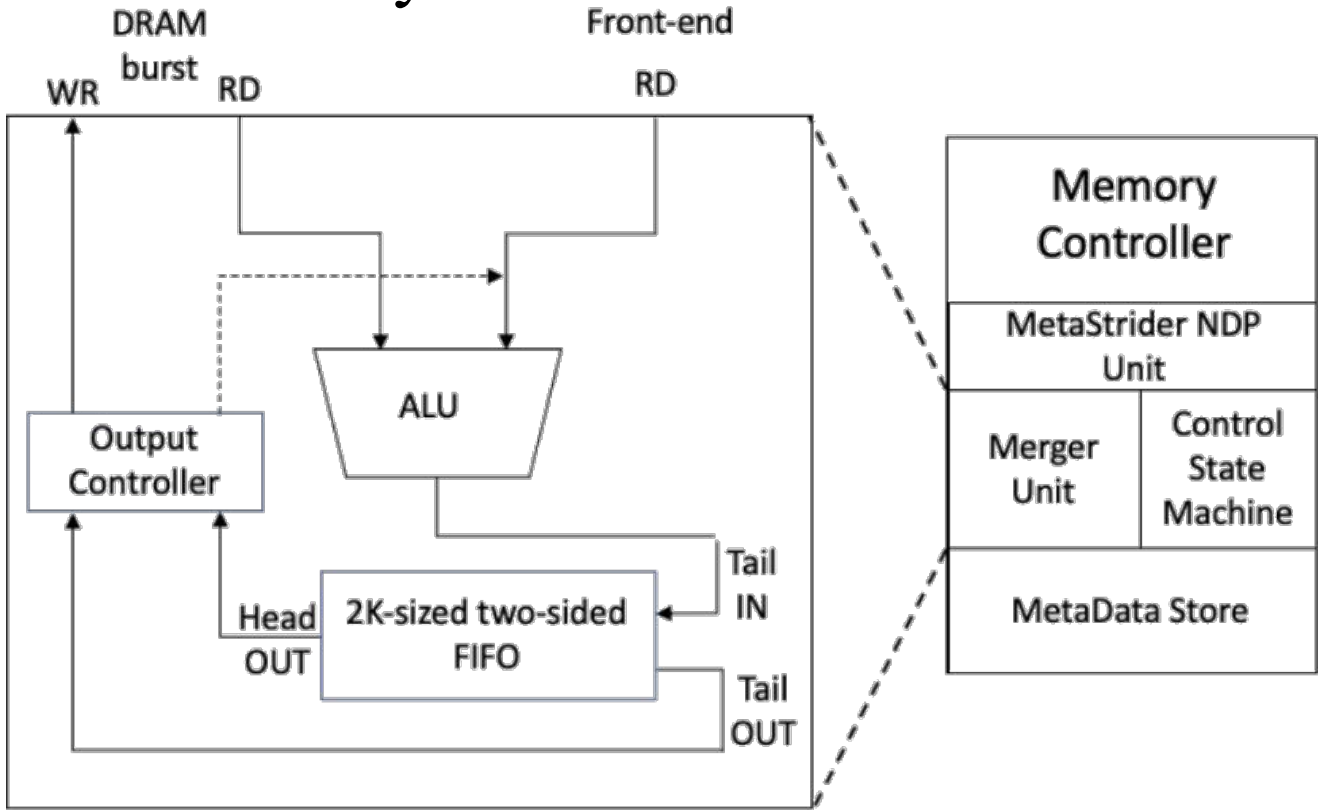
The Lucata / Emu architecture focuses on fine-grained memory access.

This really exists. And is PGAS. Because...



- No cache.
- The OS is handled by the “boring” part.
- Physically distributed memory.
- Many threads to tolerate…
- LOCAL LATENCIES.
  - Read remotely? MIGRATE.
  - Small context, one flit.
  - Plenty of references.
- Oh, and by the way…
  - Narrow channel DRAM: No wasting cache lines (so not using ⅛ BW).
  - Memory-side processing.
  - Including floating-point accumulation.

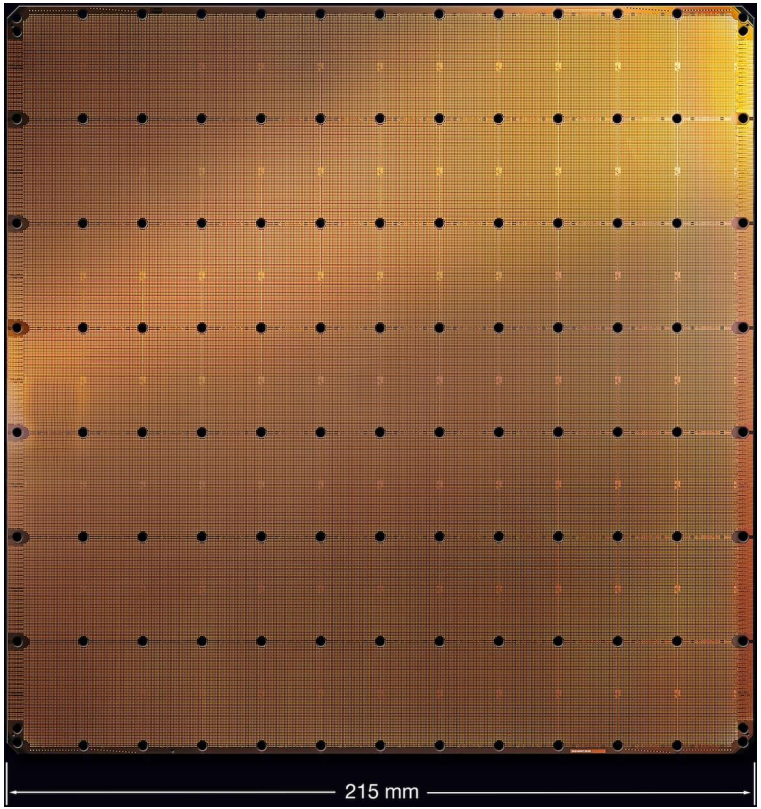
Not the only idea out there.



- **Metastrider**
- Maybe embed sparse gathers in memory (CAMS)...
- 5.3x energy savings
- 11% performance boost

Sriseshan Srikanth, Anirudh Jain, Joseph M. Lennon, Thomas M. Conte, Erik Debenedictis, and Jeanine Cook. 2019. MetaStrider: Architectures for Scalable Memory-centric Reduction of Sparse Data Streams. *ACM Trans. Archit. Code Optim.* 16, 4, Article 35 (Janua 2020), 26 pages. DOI:https://doi.org/10.1145/3355396

Totally nuts ideas……..



Borrowed from Cerebras Systems, Inc.

What if……

- You could have a hardware dataflow architecture?
-

Totally nuts ideas……...

A Rogues Gallery photo!

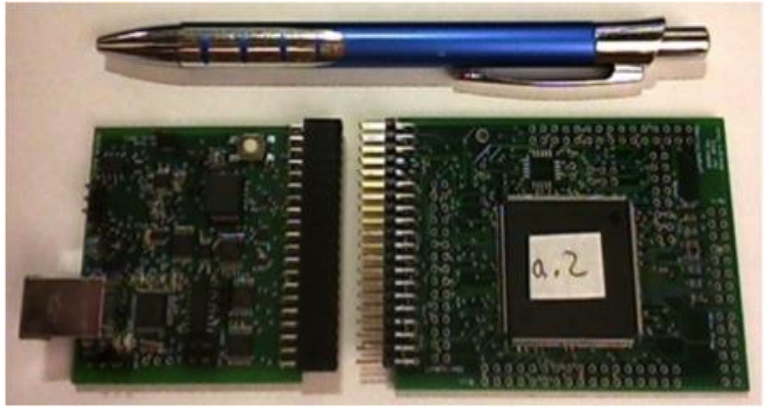
What if……

- You could have a hardware dataflow architecture?
- You could have “infinite” storage **with** logic?
-

Totally nuts ideas……..

What if……

- You could have a hardware dataflow architecture?
- You could have “infinite” storage with logic?
- You could have programmable analog devices?
  - Neuromorphic? Waiting on the recount.



A Rogues Gallery photo!

The crazy thing is that all these exist.

So how are we taking advantage?

I apologize to the non-US folks. I only know our labs with testbeds:

- DoE: ORNL, LBNL, ANL, SNL (Sandia, not Saturday Night), …
- NSF: Georgia Tech's **Rogues Gallery**, others…
- A64fx came from Japan / England.
- My preference baseline: RISC-V
  - (because you can bolt anything alongside)

No, really, go out and play! Those ideas from the 80s and before? YUP!

BTW, there are open foundries now…
No reason why algorithms folks should be quiet.

My photos are thanks to the Franco-Berkeley Fund.