Pathfinding for Post-Exascale HPC

Oklahoma Supercomputing Symposium
September 30, 2020

John Shalf
Department Head for Computer Science
Lawrence Berkeley National Laboratory

jshalf@lbl.gov
Technology Scaling Trends

Exascale in 2021... and then what?

Performance

Year

- Transistors
- Thread Performance
- Clock Frequency
- Power (watts)

Exascale Happens in 2021-2023

Figure courtesy of Kunle Olukotun, Lance Hammond, Herb Sutter, and Burton Smith
Specialization:
Natures way of Extracting More Performance in Resource Limited Environment

Powerful General Purpose

Many Lighter Weight
(post-Dennard scarcity)

Many Different Specialized
(Post-Moore Scarcity)

Xeon, Power

KNL AMD, Cavium/Marvell, GPU

Apple, Google, Amazon
Samba Nova
Extreme Hardware Specialization is Happening Now!

This trend is already well underway in broader electronics industry.

Cell phones and even megadacenters (Google TPU, Microsoft FPGAs...)

(and it will happen to HPC too... will we be ready?)

40+ different heterogeneous accelerators in Apple A11 (2019)
Large Scale Datacenters also Moving to Specialized Acceleration
The Google TPU

Deployed in Google datacenters since 2015

- “Purpose Built” actually works - Only hard to use if accelerators was designed for something else
- Could we use TPU-like ideas for HPC?
- Specialization will be necessary to meet energy-efficiency and performance requirements for the future of DOE science!

<table>
<thead>
<tr>
<th>Model</th>
<th>MHz</th>
<th>Measured Watts</th>
<th>TOPS/s</th>
<th>GOPS/s /Watt</th>
<th>GB/s</th>
<th>On-Chip Memory</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Idle Busy 8b</td>
<td></td>
<td>8b FP</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Haswell</td>
<td>2300</td>
<td>41 145 2.6 1.3</td>
<td>18 9</td>
<td>51</td>
<td>51 MiB</td>
<td></td>
</tr>
<tr>
<td>NVIDIA K80</td>
<td>560</td>
<td>24 98 -- 2.8</td>
<td>29 160</td>
<td>8 MiB</td>
<td></td>
<td></td>
</tr>
<tr>
<td>TPU</td>
<td>700</td>
<td>28 40 92 --</td>
<td>2,300</td>
<td>34</td>
<td>28 MiB</td>
<td></td>
</tr>
</tbody>
</table>

Notional exascale system:
2,300 GOPS/W \to? 288 GF/W (dp) \to a 3.5 MW Exaflop system!
AWS CEO Andy Jassy:

“AWS isn't going to wait for the tech supply chain to innovate for it and is making a statement with performance comparisons against an Intel Xeon-based instance. The EC2 team was clear that Graviton2 sends a message to vendors that they need to move faster and AWS is not going to hold back its cadence based on suppliers.”
Why does it Matter?

Why should we specialize?
HEP: Computing challenges for Particle Tracking

Exponential growth of the current ATLAS Inner Detector reconstruction time with increased luminosity ...

... but computing power no longer increasing at exponential rates!

New approaches must be developed to satisfy growing computing demands of the experiment
Impacts of Moore’s Law Tapering on DOE Science

Jim Siegrist May 2018 presentation to HEP Advisory Panel

- Computing capacity for LHC-II off by $850M compared to original estimates

- A major factor in mis-projection was due to earlier assumption that Moore’s Law would continue unabated

**HEP Computing Strategy**

- Successful implementation of the broad science program envisioned by P5 will require an equally broad and foresighted approach to the computing challenges
  - Meeting these challenges will require us to work together to more effectively share resources (hardware, software, and expertise) and appropriately integrate commercial computing and HPC advances

- Last year OHEP stood up an internal working group charged with:
  - Developing and maintaining an HEP Computing Resource Management Strategy, and
  - Recommending actions to implement the strategy

- Working group began by conducting an initial survey of the computing needs from each of the three physics Frontiers, and assembled this into a preliminary model

- Energy Frontier portion alone was a large factor beyond the current computing budget
  - Large data volumes with the HL-LHC require correspondingly large amounts of computing to analyze it
    - Grid-only solution: **$850M ± 200M**
    - Using the experiments’ estimates of future HPC use reduces this to **$650M ± 150M**
Mission Need doesn’t end with Exascale

**HENP:** compute requirements grow exponentially relative to luminosity

**BES Light Sources & CryoEM:** Double-exponential growth of camera data rates (100k FPS)

**Cloud-Resolving Climate Models:** Kilometer scale climate models still out of reach (~1 SYD in 2010, ~5 SYD in 2020)

What if we are successful in creating AI driven (no-human in the loop) experiments? What kind of data processing would be needed to keep up with that?
Architecture Specialization for Science

(hardware is design around the algorithms) can’t design effective hardware without applied math

But what are the right specializations to include? What is the cost model (we know we cannot afford to spin our own chips from scratch)

What is the right partnership/economic model for the future of HPC?

The role government research is to understand these trade-offs.
Specialization
Purpose built machines for big science targets.
Example: Google TPU. For DOE, DFT is 25% of workload

Heterogeneous Integration
Co-integration of many heterogeneous accelerators

Example: Apple Bionic chip, AWS Graviton2, Project38.

Resource Disaggregation
Photonic MCMs to enable reconfigurable nodes/systems
Example: Facebook/Google. Just DRAM utilization diversity in DOE could benefit from this.
Specialization

Purpose built machines for big science targets.

Example: Google TPU. For DOE, DFT is 25% of workload

Heterogeneous Integration

Co-integration of many heterogeneous accelerators

Example: Apple Bionic chip, AWS Graviton2, Project38.

Resource Disaggregation

Photonic MCMs to enable reconfigurable nodes/systems

Example: Facebook/Google. Just DRAM utilization diversity in DOE could benefit from this.
Algorithm-Driven Design of Programmable Hardware Accelerators

Example: LS3DF/Density Functional Theory (DFT)

**What:** Design the hardware acceleration around the target algorithm/application
- Purpose-built acceleration
- Science-led reference algorithm design

**Why:** Huge opportunities to improve performance density and efficiency
- FFT hardware accelerator 50x-100x faster than GPU (using SPIRAL generator)

**How:** Target Density Functional Theory
1. Large fraction of the DOE workload
2. Mature code base and algorithm
3. LS3DF formulation minimizes off-chip communication and scales O(N)
The DFT kernel for each fragment

Communication Avoiding LS3DF Formulation – Scales $O(N)$

$h(i, j) = \langle \psi_i | H | \psi_j \rangle$

Sub_diag, *
Hpsi, *
Precond. CG step
Projection, *
Line minimiz.
Orth., *

CGRA or FPGA

O($N^2 \log(N)$)
Comm bound if non-local
3D parallel FFT

TSQR & Choelesky
ZGEMM

O($N^3$)
Compute-bound

LS3DF O($N$) Algorithm Formulation
Minimizes off-chip Communication

One patch per CGRA
400 bands/patch

Compute Intensive Kernels
Targeted for HW Specialization
Von-Neumann Instruction Processors vs. Hardware Circuits

(must redesign for static dataflow and deep flow-through pipelines)

**FPGA** (*Field Programmable Gate Array*): Granularity of these operations and wires are single bits

**CGRA** (*Coarse Grain Reconfigurable Array*): Programmability & ALUs at word granularity improves speed and density!! *(Cerebras, GraphCore, SambaNova, LPU)*

**ASIC or Chiplet** (*custom circuit*): Another factor of 10x on density and energy efficiency.
Algorithm Reformulated as Custom Circuit

Von Neumann CPU

DRAM

GEMM

iFFT3D

FFT3D

Point wise

Control Unit

Arithmetic Logic Unit

Dataflow (FPGA, GraphCore etc.)

Data Flows In

Results Flow Out

See Also Torsten Hoefler “StreamBLAS” for FPGA
Preliminary Performance on CGRA HΨ

**Eigenvalue Problem**

\[ h(i,j) = \langle \psi_i | H | \psi_j \rangle \]

\[ P_i = H \psi_i - e_i \psi_i \]

\[ P_i = A \left( P_i - \frac{\lambda_i}{\lambda_0} \psi_i \right) \]

\[ P_i = P_i - \sum_{j=1,i}^{n} \langle \psi_i | \psi_j \rangle \]

\[ \psi_i = \psi_i \cos \theta_i + P_i \sin \theta_i \]

\[ h(i,j) = \langle \psi_i | H | \psi_j \rangle \]

**Von Neumann CPU or GPU**

```c
int main()
    int n = 0;
    while (n < 300)
        n = n + 5;
        printf("n = %d\n", n);
        pause(200);
    if (n == 50) break;
    printf("All done!\n");
```

**Dataflow Algorithm Reformulation**

**Mapping onto Custom Hardware**

**Results**

<table>
<thead>
<tr>
<th>Platform</th>
<th>Time for Contraction</th>
<th>Speedup over CPU</th>
<th>Speedup over GPU</th>
</tr>
</thead>
<tbody>
<tr>
<td>CPU (Haswell/Cori Phase 1) node</td>
<td>1.375</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>GPU (NVIDIA 1080)</td>
<td>0.5</td>
<td>2.75</td>
<td>1</td>
</tr>
<tr>
<td>CGRA unoptimized</td>
<td>0.23</td>
<td>6</td>
<td>2.2</td>
</tr>
<tr>
<td>CGRA optimized</td>
<td>0.023</td>
<td>60</td>
<td>21.7</td>
</tr>
</tbody>
</table>

Delivered Speedups (compared to optimized code) of “custom” DFT accelerator running on CGRA

Thom Popovici, Andrew Canning (FFTx), Zhengji Zhang (NERSC), Franz Francetti (CMU/FFTx)
Heterogeneous Integration
Specialization
Purpose built machines for big science targets.

Example: Google TPU. For DOE, DFT is 25% of workload

Heterogeneous Integration
Co-integration of many heterogeneous accelerators

Example: Apple Bionic chip, AWS Graviton2, Project38.

Resource Disaggregation
Photonic MCMs to enable reconfigurable nodes/systems

Example: Facebook/Google. Just DRAM utilization diversity in DOE could benefit from this.
Project 38 (P38) is a set of vendor-agnostic architectural explorations involving DOD, the DOE Office of Science, and NNSA

- **Near-term goal:** Quantify the performance value and identify the potential costs of specific architectural concepts against a limited set of applications of interest to both the DOE and DOD.
- **Long-term goal:** Develop an enduring capability for DOE and DOD to jointly explore architectural innovations and quantify their value.
- **Stretch goal:** Specification of a shared, purpose built architecture to drive future DOE-DOD collaborations and investments. (purpose-built HPC by 2025)

**Accomplishments**

- Released initial project report through NITRD in 2020 that identifies 8 promising architecture enhancements that can significantly improve application performance.
- Working with Arm, AMD (LBL/ANL/PNNL), and Micron (Sandia/LLNL) to assess feasibility and develop cost models.
- ANL evaluating impact of diverse specializations on the programming environment & compiler technologies.

**Phase 1 Report:** [Link](https://www.nitrd.gov/Presentations/files/HPC-Performance-Improvements-Project-38.pdf)

**Related Effort at LANL**

Jason Pruett

“Tailored Computing”

(whitepaper forthcoming)
Recapping Key P38 Technology Explorations

• Fixed Function Accelerators & COTS IP (*Extreme Heterogeneity*)
  — RISC-V and ARM cores
  — Fixed function FFT (Generated by SPIRAL)

• Word Granularity Scratchpad Memory (Gather Scatter):
  — Gather-scatter within processor tile
  — more effective SIMD

• Recoding engine (Efficient programmable FSM & data reorg.)
  — Sub-word granularity and high control irregularity
  — Handles branch-heavy code (avg. 20x improvement over processor core)
  — One lane is 1/100th the size of a x86 processor core

• Hardware Message Queues (Lightweight Interprocessor Communication)
  — Gather-scatter between processor tiles
  — Async between tiles to eliminate overhead of barriers
• What if HPC adopted Smartphone SoC Strategy -- *mix fixed-function accelerators with programmable cores*

• Target commonly used scientific primitives/libraries
  – BLAS (level 1, 2, 3)
  – FFT (FFTW or SPIRAL interface)
**FFT Example With FFTx (Francetti, Popovic, Canning)**

For FFT of size $N$
- Storage = $N \times$ operand_size
- Compute = $5/2 \times N \times \log_2(N)$ FLOPs
- Use Pseudo-2D algorithm for large FFTs

**Single FFT Accelerator Resource**
- **Assumptions: Spiral HW Generator**
  - 1GHz @ 14nm technology node
  - 2M point transform (data off-chip)
  - HPC Challenge Benchmark: Single precision (Float32) complex, out-of-place

- **Limit: 100 GB/s off-chip memory**
  - 16k points on-chip engine
  - Analytic model for FP limit $\sim 1.5$TFLOPs SP
  - $4.5 \text{mm}^2$ area for compute @ 14nm

- **Limit: 1TB/s off-chip memory**
  - $\sim 10k$ MADD + $\sim 5k$ add -> $15k$ FP@1GHz
  - Analytical model for FP limit $\sim 15$TFLOPs SP
  - $47 \text{mm}^2$ area for compute @14nm
IP Reuse is Key

This is the *real* power of the ARM ecosystem (it's not just about Arm cores or Cavium)

- Leverage commodity ecosystems
- Get commercially supported IP where there is a market to support it
- Use open-source IP where the government needs to develop technology to serve its needs
- Partner with system integrators & chip vendors for realization of systems
  (new sustainable economic model for HPC)

Average % of reused IP Blocks
- 2016 average:
  • 175 IP blocks
  • 80% reuse

Average # of IP Blocks
- 18 blocks
- 55 blocks
- 110 blocks
- 52 blocks

SEMICO Research Corporation, 2014
Resource Disaggregation
Specialization
Purpose built machines for big science targets.

Example: Google TPU. For DOE, DFT is 25% of workload

Heterogeneous Integration
Co-integration of many heterogeneous accelerators

Example: Apple Bionic chip, AWS Graviton2, Project38.

Resource Disaggregation
Photonic MCMs to enable reconfigurable nodes/systems

Example: Facebook/Google. Just DRAM utilization diversity in DOE could benefit from this.
Diverse Node Configurations for Datacenter Workloads

**Training**
- 8 connections: GPU
- 8 links to HBM (weights)
- 8 links: to NVRAM
- 1 links: to CPU (control)

**Inference**
- 16 links to TOR (streaming data)
- 8 links HBM (weights)
- 1 link: CPU

**Data Mining**
- 6-links: HBM
- 15 links: NVRAM (capacity)
- 4 links: CPU (branchy code)

**Graph Analytics**
- 16 links HBM
- 8 links TOR
- 1 Link CPU
Memory Disaggregation

About 15% of NERSC workload uses more than 75% of the available memory per node.

And ~25% uses more than 50% of available memory.

But 75% of Haswell job hours (60% of KNL) use < 25% memory

Overestimate: maxrss $\times$ ranks_per_node
Assumes memory balance across MPI ranks.
Most solutions current disaggregation solutions use Interconnect bandwidth (1 – 10 GB/s)
But this is significantly inferior to RAM bandwidth (100 GB/s – 1 TB/s)
Interposers are the right point of intersection where copper pin bandwidth density could match photonics bandwidth density!

- **Good News:** Extend Bandwidth Density and lower power/bit
- **Bad News:** Limited (~2cm) reach — Cannot get outside of the package (*but we need to!!*)

**Attachment technologies**
- **Solder**
- **Copper-copper**
- **Hybrid bonding** (Ziptronix DBI)

**Today typically 400 \text{pm}** pitch ~ 250 \text{pm} pitch demonstrated ~ Potential for 50 \text{pm} pitch ~ Potential for sub-10 \text{pm} pitch — Enabled by sub-10 \text{pm} alignment tools

IBM Ziptronix
Impedance Matching to our Packaging Technology

In-package integration
Solder Microbumps & Copper Pillars @ 10Gbps
Wide and Slow!

DWDM Using Silicon Photonics
Ring Resonators @ 10 Gigabits/sec per chan
Many channels to get bandwidth density
Wide and Slow!

Comb Laser Sources
Single laser to efficiently generate 100s of frequencies
Wide and Slow!
Photonic MCM (Multi-Chip Module)

Comb Laser Source with DWDM Silicon Photonics
Wide-and Slow for high speed links

Photonic SiP

Silicon interposer

Package substrate

1024 data links / HBM stack @ 500MHz
Photonic MCM (Multi-Chip Module)

- Fiber carrying 0.5 - 1 Tb/s
- High-Density fiber coupling array with 24 fibers = 6-12 Tb/s bidirectional = 0.75 – 1.5 TB/s
- Silicon interposer
- Package substrate
- Fiber coupler pitch: 10s of um
- High-Density fiber coupling array
- Photonic SiP
- Optical switch
- To other nodes
### Training
- 8 connections: Peer GPU
- 8 links to HBM (weights)
- 8 links to NVRAM
- 1 link to CPU (control)

### Inference
- 16 links to TOR (streaming data)
- 8 links HBM (weights)
- 1 link: CPU

### Data Mining
- 6 links: HBM
- 15 links: NVRAM (capacity)
- 4 links: CPU (branchy code)

### Graph Analytics
- 16 links HBM
- 8 links TOR
- 1 Link CPU

**Configure for Training**

**Configure for Inference**
PINE: Photonic Integrated Networked Energy Efficient Datacenters

Resource Disaggregation to custom-assemble diverse accelerators for diverse workload requirements

1) Energy-bandwidth optimized optical links

2) Embedded silicon photonics into OC-MCMs

3) Bandwidth steering for Custom Node Connectivity

Optically Interconnectivity for Deep Disaggregation
MCM can be reconfigured to accelerate different applications

1 Tb/second per fiber

ENLITENED
Economic Models
3. The Decline of Computers as a General Purpose Technology

Papers
1. The Economic Impact of Moore’s Law
2. There’s Plenty of Room at the Top: What will drive computer performance after Moore’s Law?
3. The Decline of Computers as a General Purpose Technology

The Top

<table>
<thead>
<tr>
<th>Technology</th>
<th>Software</th>
<th>Algorithms</th>
<th>Hardware architecture</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>01010011 01100011</td>
<td>New algorithms</td>
<td>Hardware streamlining</td>
</tr>
<tr>
<td></td>
<td>01101001 01100011</td>
<td>New problem domains</td>
<td>Processor simplification</td>
</tr>
<tr>
<td>Opportunity</td>
<td>Software performance</td>
<td>Tailoring software</td>
<td>Domain specialization</td>
</tr>
<tr>
<td></td>
<td>engineering</td>
<td>to hardware features</td>
<td></td>
</tr>
<tr>
<td>Examples</td>
<td>Removing software</td>
<td>New machine models</td>
<td></td>
</tr>
<tr>
<td></td>
<td>bloating</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The Bottom

- Technology
- Opportunity
- Examples

Papers
1. The Economic Impact of Moore’s Law
2. There’s Plenty of Room at the Top: What will drive computer performance after Moore’s Law?
3. The Decline of Computers as a General Purpose Technology
IP Reuse is Key: \textit{(IP is the commodity & cost driver)}

Neil Thompson

- **2016 average:**
  - 175 IP blocks
  - 80% reuse

![Graph showing average number of reused IP blocks over time]

- **Average # of IP Blocks**
- **Percent of Reuse**
- **Avr. Number of IP Blocks**

SEMICO Research Corporation, 2014
Chiplets and Wafer-Scale Integration as path for Heterogeneous Integration

CHIPS modularity targets the enabling of a wide range of custom solutions
Industry: Heterogeneous Integration Roadmap

2019 Edition

http://eps.ieee.org/hir

HPC and Megadatacenters is 2nd chapter
In the era of the "universal computer" scale was the correct answer to deliver value to our scientific customers. In this post-moore/post-exascale era, that is not a viable approach to continuing to deliver value to our customers. It isn’t scale, it must be differentiation and targeted specialization. Scale demanded we focus on capital costs. The new era must increase focus on development costs to meet the demands of science. The "cloud" does not mitigate this outcome.
Project 38 -- Background

**DOD and DOE recognize the imperative to develop new mechanisms for engagement with the vendor community, particularly on architectural innovations with strategic value to USG HPC.**

Project 38 (P38) is a set of vendor-agnostic architectural explorations involving DOD, the DOE Office of Science, and NNSA (these latter two organizations are referred to in this document as “DOE’’). These explorations should accomplish the following:

- **Near-term goal:** Quantify the performance value and identify the potential costs of specific architectural concepts against a limited set of applications of interest to both the DOE and DOD.

- **Long-term goal:** Develop an enduring capability for DOE and DOD to jointly explore architectural innovations and quantify their value.

- **Stretch goal:** Specification of a shared, purpose built architecture to drive future DOE-DOD collaborations and investments. (purpose-built HPC by 2025)