FACULTY
OF INFORMATICS
Masaryk University

PA039: Supercomputer Architecture and Intensive Computing

Myth and Legends in HPC

Luděk Matyska

Spring 2024

The Source

S. Matsuoka, J. Domke, M. Wahib, A. Drozd, and T. Hoefler (RIKEN & ETH): *Myth and Legneds in High-Performance Computing*, 25.11.2023
`https://arxiv.org/pdf/2301.02432`

See also the presentation of E. Laure at e-INFRA CZ conference in 2024 (in IS)

12 Myths

1. Quantum Computing Will Take Over HPC
2. Everything Will Be Deep Learning
3. Extreme Specialization as Seen in Smartphones Will Push Supercomputers Beyond Moore’s Law
4. Everything Will Run on Some Accelerator
5. Reconfigurable Hardware Will Give 100X Speedup
6. We will Soon Run at Zettascale
7. Next Generation Systems Need More Memory per Core
8. Everything Will Be Disaggregated
9. Application Continue to Improve, Even on Stagnating Hardware
10. Fortran Is Dead, Long Live the DSL
11. HPC Will Pivot to Low or Mixed Precision
12. All HPC Will Be Subsumed by the Clouds

1. Quantum Computing Will Take Over HPC

- Quantum ”supremacy” needs exponential speedup compared to classical algorithms
  - Shor’s algorithm exponential
  - Grover’s algorithm quadratic
- Memory bandwidth
  - Near future Quantum hardware – Gigabits/s
  - Contemporary single-chips – Terabits/s
- ”Cooldown” and input reset delay
- Hybrid systems may be useful, but they need strong classical HPC part

2. Everything Will Be Deep Learning

- Similar to QC hype/myth
- Enormous success with generative (language) models
  - However, success based on the availability of more ”classical” HPC resources
- Uncertainty quantification and explainability are problems
  - How to find out when the AI model starts ”hallucinating”
- AI models versus first-principle simulations
  - AI clear winner where the first-principle models are not available
  - But a large number needs rigorous accuracy and precision guarantees
- Shat will be the new areas with deep AI impact?

3. Extreme Specialization as Seen in Smartphones Will Push Supercomputers Beyond Moore's Law

- System on CHipo(SoC) extremely successful in smartphones and similar devices
- Could future supercomputers be built using the same principles?
  - Extensive hardware customization to fit all facets of a HPC workload
- The only really successful accelerator in HPC is the GPU (which became the GPGPU)
  - High memory bandwidth
  - Currently adopted by e.g. Intel Saphire Rapids with HBM
- Reasons against this myth
  - Strong (as per SoC) versus weak (as per HPC success) scaling
  - No more dark silicon available in the system
  - Software and productivity

4. Everything Will Run on Some Accelerator

- See the previous slide/myth
- Standard workload types
  - C Compute-bound
  - B Memory bandwidth-bound
  - L Memory latency-bound
- FPGA – Field Programmable Gate Array
- CGRA – Coarse-Grained Reconfigurable Array
- Insufficient experience with more than CPU/GPU combination

5. Reconfigurable Hardware Will Give 100X Speedup

- A follow-up of the previous one – FPGA potential
- In the past 20 years, we had seen rise and fall of the reconfigurable hardware
  - Never got sufficient support at applications' level
  - Lower energy efficiency

6. We will Soon Run at Zettascale

- What is Zettascale?
  - *Zettaflop system* – any computer with more than **10^{2}1** double-precision floating point operations per second
  - *Zettaop system* – any computer theoretically capable of **10^{2}1** operations per second
  - *Zettascale system* – any computer capable to execute an application with a performance of over 1 zettaflop/s in fp64
- Extrapolation
  - 1.068 teraflop/s, 0,85 MW – summer 1997, ASCI Red
  - 1.026 petaflop/s, 2,35 MW – summer 2008, Roadrunner
  - 1.05 exaflop/s, 35 MW – spring 2021, OceanLight (China, unofficial)
  - 1.1 exaflop/s, 21.1 MW – summer 2022, Frontier

Three orders of magnitude per 10+ years

6. We will Soon Run at Zettascale

- Can we continue?
  - 50-100 MW
  - Need to increase energy efficiency of fp64 operations 200-350x
  - The interconnect complexity – up to 2.5 GW
  - *Zettaop system* more feasible, with 50 MW ”only”
- A perspective
  - *Zettaflops system* in 2037 with 200 MW
  - *Zettascale system* in 2038

7. Next Generation Systems Need More Memory per Core

- A simplistic statement
- The need for large memory may be a legacy
- A balanced approach is needed, not emphasising just one component/feature

8. Everything Will Be Disaggregated

- Silicon Photonic and all-optical interconnects
  - High bandwidth density significantly reduces the number of I/O lanes
  - Power consumptions and crosstalk are distance neutral
  - Propagation loss is low
- Obstacles
  - low-cost manufacturing
  - optical switching
- Latency – speed of light for HPC use

9. Application Continue to Improve, Even on Stagnating Hardware

- No more Mooore’s law dependency
- Post-Moore era performance road:
    - Architectural innovations
    - Alternative materials and technologies
    - Abandoning the von-Neumann paradigm
- ”Algorithmic Moore’s law”
    - Can provide exponential improvement
    - Has it’s own limits: numerical stability, asymptotic limits, …

10. Fortran Is Dead, Long Live the DSL

- A more general question on the value of general-purpose programming languages
- Where should the portability layer be located?
- If we need this generality, than Fortran evolution will continue

11. HPC Will Pivot to Low or Mixed Precision

- Perhaps, but we have to see how far the mixed precision arithmetic can lead us

12. All HPC Will Be Subsumed by the Clouds

- Cloudification of supercomputers
- Supercomputerization of clouds