Widespread AI Adoption Creates New Demands

Tom Sorensen

May 2024

New Category of Advanced Computing

- HPC users adopting/integrating AI at high rates
- LLMs received most attention among a diverse field of model and method types
- ~90% of HPC users surveyed currently or plan to use AI methods for workloads
- AI methods introduce new demands on sites:
  - Hardware (processors, interconnect, data access)
  - Software (data management, queueing, dev tools)
  - Expertise (procurement strategy, maintenance, troubleshooting)
  - Regulatory (data provenance, privacy, legal)

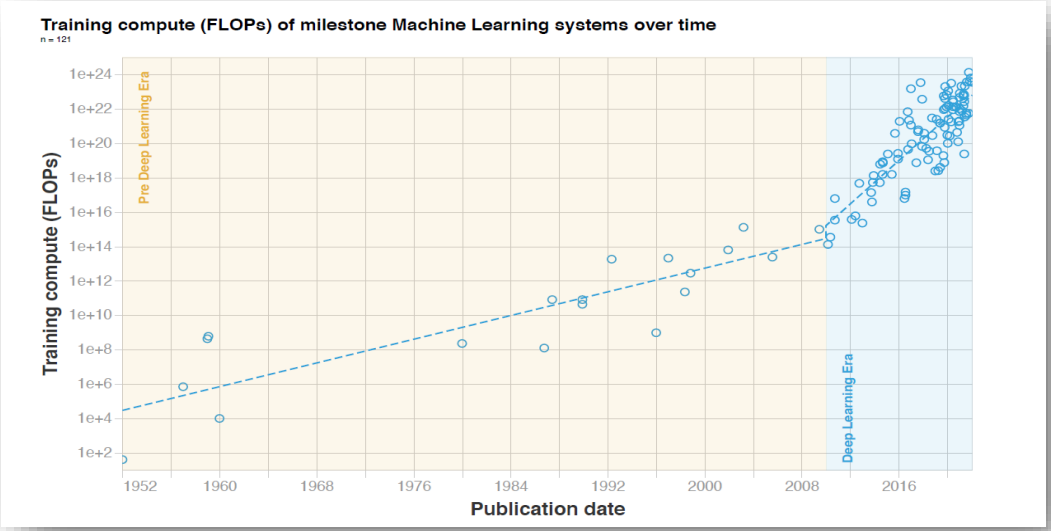
Framing LLM/HPC Requirements

Key elements dominate scaling of LLMs on HPCs

- **Compute: the absolute number of floating-point operations needed to train an LLM to a desired degree of accuracy**
- **Dataset size: input dataset used for training the LLM**
- **Model size: number of tokens or parameters**
  - The larger the number of parameters, the more nuance in the model's understanding of each word's meaning and context
- **Integration: Ability to introduce effective models into a production or experimental environment**
  - Users identify ease of AI integration as the highest consideration when adopting AI tools into production environments
- **LLMs require unique HPC capabilities**

LLMs Consume Significant Flops

LLM flops growth eclipses Top 500 growth



- Pre 2010:
  - On the order of 2 X 10^{12} (200 Tflops)
  - Flops requirements doubling every 21.3 months
  - But not a lot of data points
- Post 2010 to Current:
  - Currently on the order of 6 X 10^{22} flops (60 Zettaflops)
  - Flops requirements doubling every 5.6 months
  - Roughly 11X faster than HPC Top 1 Linpack performance growth rate

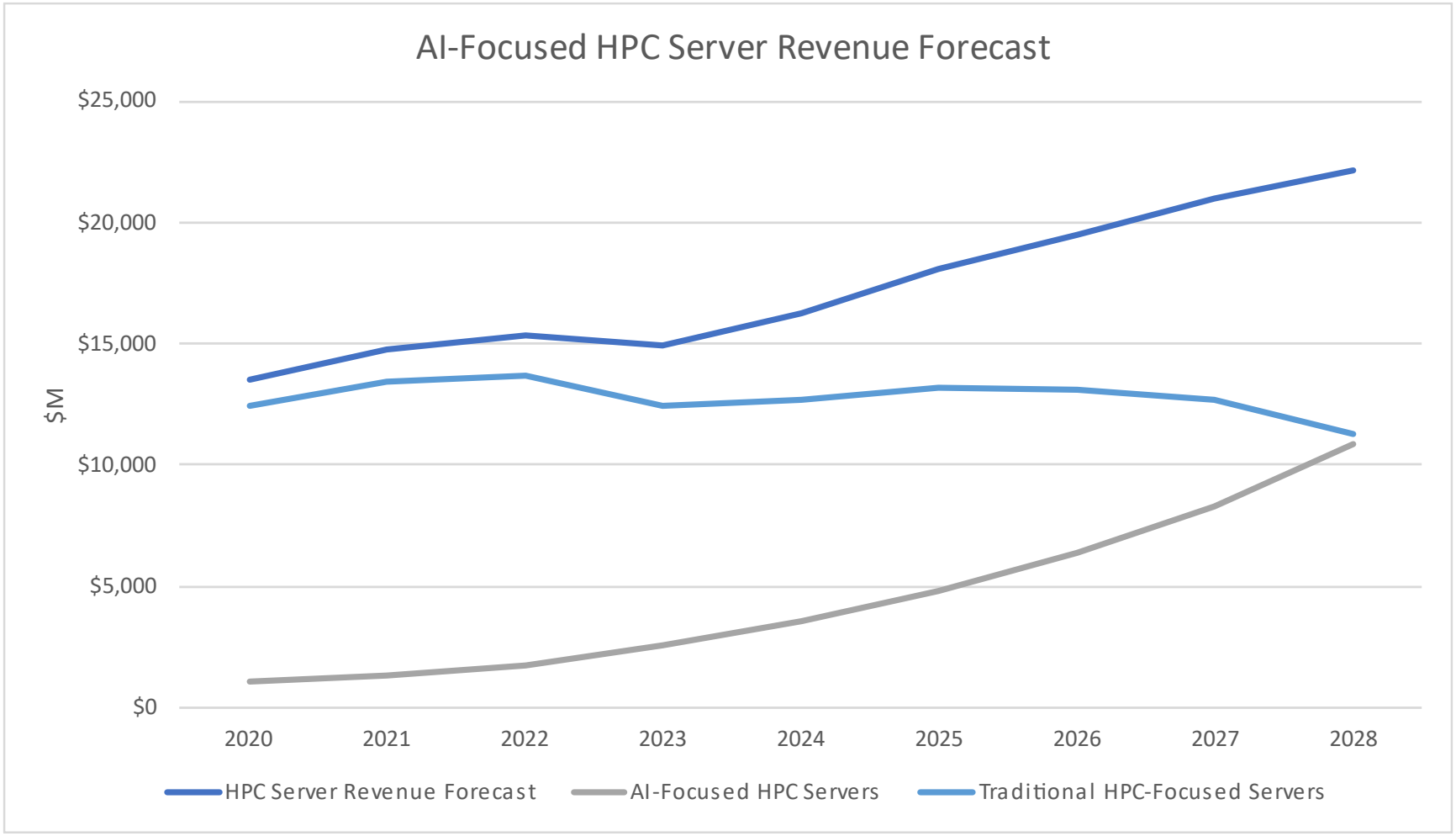
See Compute Trends Across Three Eras of Machine Learning, arXiv:2202.05924

Select Key Findings from AI Use Study

From a survey of HPC users leveraging LLMs

- **1:** LLMs are considered to be an important productivity enhancer asset for both current and planned HPC-related activity in workload management and data analysis, but not necessarily as a computational accelerator.
- **3:** Respondent organizations are exploring a broad set of LLM-related end uses.
- **5:** Adoption may be limited due to the high cost of hardware such as GPU and other accelerators.
- **9**: Open source is currently the most preferred option for accessing LLM software.
- **10:** Survey respondents looked to a wide range of LLM expertise to support the various stages of LLM development spanning foundation model construction, fine-tuning procedures, LLM integration into existing workloads, and supporting inference operations.

AI Server Forecast ($M)



Putting This All Together

Is this (another) new HPC architectural paradigm in the works?

- **Based on a recent LLM analysis by Riken**
- **GPT variant flops requirements**
  - GPT-3.5 (ChatGPT): $3X10^{24}$ flops (estimated)
  - GPT-4.0: $3X10^{25}$ flops (estimated)
- **OpenAI System: Microsoft/Open AI collaboration**
  - Top 5 system when stood up
  - GPU-based BF16 312 Tflop/s x 25,000 = 7.8 Eflop/s TPP
  - GPT-3.5 (ChatGPT): 4.5 days X 2
  - GPT-4.0: 45 days X 2
- **Fugaku:**
  - FP32 6.76 Tflop/s X 158,976 = 1.07 Eflop/s (TPP)
  - GPT-3.5 (ChatGPT): 32 days X 10
  - GPT-4.0 45 days X 2: 328 days X 10 ~= 8.9 years

Distributed Training of Large Language Models on Fugaku, https://t.co/idofa7Tjyu

HYPERION RESEARCH

QUESTIONS?

tsorensen@hyperionres.com
ejoseph@hyperionres.com