Positioning Tools and the Tools Community to Support AI/ML

Working Group Outbrief

Scalable Tools Workshop 2025

Challenge

Some top level observations:

- Our tools community has an incredibly depth of expertise, experience, and success in helping the HPC community make applications faster and more efficient.
- The AI/ML world is rapidly growing applications in both number and scale. The cost and efficiency of these applications is a growing and urgent concern.

How do we:

- Better understand the technical challenges that AI/ML computations have beyond what we know in HPC?
- Make the AI/ML world aware of our capabilities and expertise?
- Find funding opportunities at a national and international scale to move this forward?

Funding Models for Tools

- Common
  - Unbroken string of research grants
  - Direct national laboratory contracts
  - Subscription pricing
- Emerging
  - AMS Images
- Opportunities
  - Consortia

Growing the Pie for Tools

- Must explore opportunities for leveraging the explosive growth in AI/ML to grow support for sophisticated tools

How can we increase the appeal of our work to AI/ML?

Better understand the needs of the AI/ML community

- Identifying high-visibility ML users and applications
- Addressing Python/PyTorch integration challenges
- Monitoring workflow optimization and data movement

Who can we target with AI/ML tooling?

- infrastructure teams
- end users

Understand the best opportunities for impact

- How can we map HPC tools expertise to ML field needs
- Explore how we can build tools that match user abstraction levels

Strategic Focus Areas

- Identifying high-visibility ML users and applications
- Translating HPC tools expertise to needs in ML field
- Addressing Python/PyTorch integration challenges

  Developing models of performance data relevant to AI/ML applications:

  Mapping across AI model ↔ interpreted execution ↔ device
- Focusing on workflow optimization and data movement

A Challenge: Diversity of AI/ML Platforms

- Broad range of architectures are AI relevant
  - GPUs
  - TPUs
  - Dataflow architectures
    - Cerebras
    - Sambanova
- Some are less accessible to the broad tools community, e.g. TPU
- Some require very different models of measurement and analysis
  - Cerebras
  - Sambanova
- What should be the focus?

AI Needs

Concerns

- Exploiting parallelism
- Exploiting resources (compute, network, storage)
- Understanding how well computation and data movement are partitioned and mapped

Relate performance back to semantic information in AI models

- E.g., tensor sizes, data movement

Things we need to understand

- How well are AI tools working
- How efficient is the AI model

Opportunity: Integration of observability

Technical Requirements for AI/ML Tools

- Need full Python/PyTorch integration via context managers
- Must support multiple backends (TPUs, DPUs, custom accelerators)
- Require data-centric performance analysis capabilities
- Need to handle complex ML workflows with multiple components:
  - In-memory databases
  - Training processes
  - Data staging
  - Simulation code

Marketing the Capabilities of the Tools Community

- Educate federal government about benefits of investment in our area at the highest level possible
  - Beyond an individual program manager
  - Ideas for how?
    - Collect success stories of tools at scale that have yielded significant results and communicate them to highest levels
    - Try to quantify impact in some way?
- Explore options for support beyond federal government by engaging
  - Select national laboratories
  - Consortia
  - End users, especially large corporations

Potential Next Steps

- Identify proxy applications for us to use
- Online 1-day workshop with stakeholders in large-scale AI/ML
  - Have them describe what they are doing
  - Have them describe what information they need about the performance and opportunities of their computations
  - Have them describe
    - Their experiences with existing tools
      - What they've tried, what information they've gotten for their efforts
      - Things they want that they can't get
    - Their wish list for the information and insights they would like
- Encourage a DOE ASCR or NSF workshop on tools for HPC and AI/ML
  - Put out a call for white papers about ideas for new work
  - Hold workshop
  - Develop report that can be a blueprint for future funding

Existing AI Tool Publications

https://www.usenix.org/conference/atc20/presentation/zhu-hongyu

https://doi.org/10.48550/arXiv.2110.10802

https://doi.org/10.1145/3544497.3544501

https://doi.org/10.1145/3379337.3415890

https://proceedings.mlsys.org/paper_files/paper/2022/hash/b422680f3db0986ddd7f8f126baaf0fa-Abstract.html

https://doi.org/10.1109/TVCG.2023.3243228