PERF PORTABILITY VENDOR PANEL

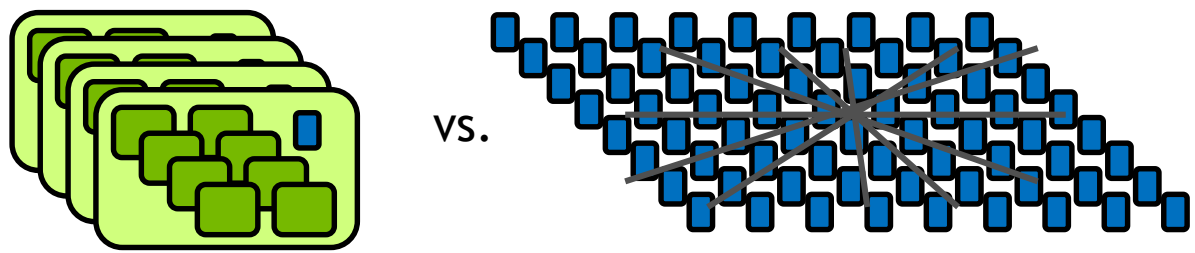
CJ Newburn, Principal HPC Architect, Compute SW

NVIDIA

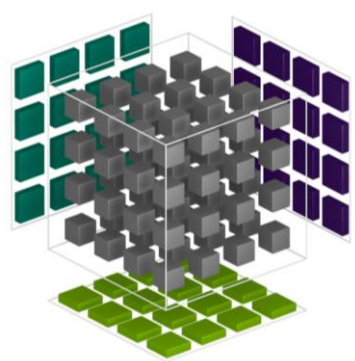
SPECIALIZATION

Use the right tool for the right job

- Absolute performance: throughput and latency cores
- Performance/Watt: dense nodes



- Deep learning
- Graphics and video

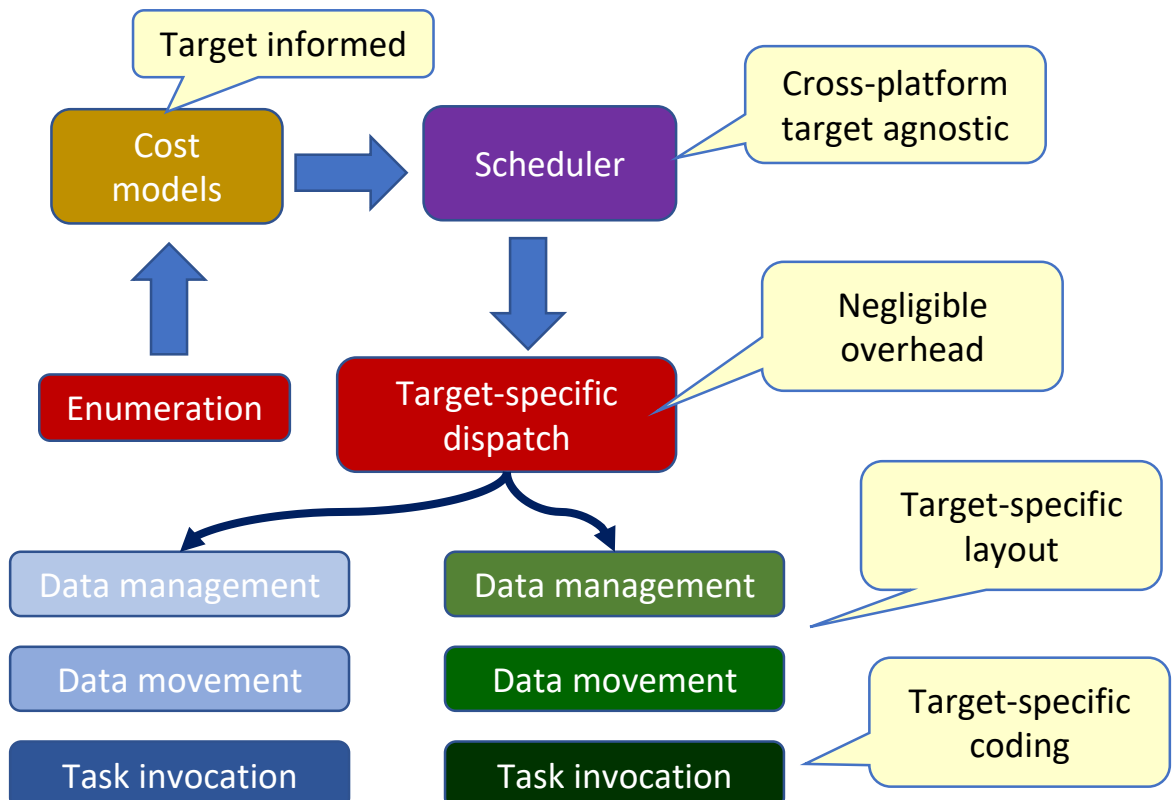


INFRASTRUCTURE FOR RETARGETABILITY

Support diversity within a common software architecture

- Common software architecture
- Common scheduler
- Common set of primitives

- Allow for diversity and specialization by managing complexity
- Enable per-target implementations and data layouts



CONVERGENCE

Synergy, new usage models

- Cloud
  - Easy deployment, provisioning
  - Apps as a service, downloaded as **containers** from a registry and orchestrated
- HPC
  - Makes good use of dense nodes and hetero resources with **retargetable frameworks**
  - Broadens and refines **programming models**, pushes **language standards**
  - Pushes absolute perf and perf/W
- AI
  - Highly-tuned **libraries** and frameworks, fast collectives

TAXONOMY

Organizing HPC + AI Convergence

Transformation

HPC + AI couple simulation with live data in real time detection/control system

Experimental/simulated data is used to train a NN that is used to for detection/control of an experiment or clinical delivery system in real time. The NN is improved continuously as new simulated / live data is acquired

Augmentation

HPC + AI combined to improve simulation time to science > orders of magnitude

Experimental/simulated data is used to train a NN that is used to replace all or significant runtime portions of a conventional simulation. The NN is improved continuously as new simulated / live data is acquired

Modulation

HPC + AI combined to reduce the number of runs needed for a parameter sweep

Experimental/simulated data used to train a NN which steers simulation/experiment btwn runs

The steering NN can be trained continuously as new simulated / live data is acquired

Potential for Breakthroughs in Scientific Insight

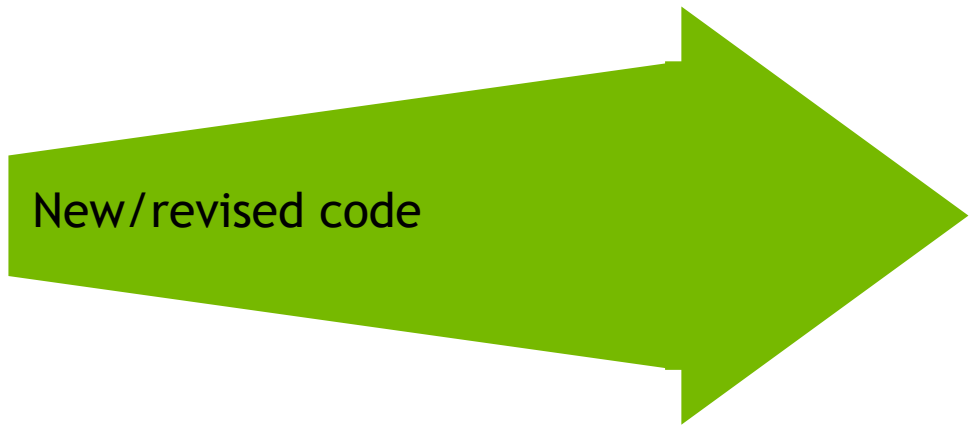
PROVIDING ACCESS TO PERFORMANCE

Meeting our customers where they are, offering a path forward

- % lines of code gains, ROI

Aggressive tuning for the target platform

- Exposing maximal parallelism
- Extreme scaling
- Tailored abstractions



Limited effort

- Standard interfaces
  - C++, OpenMP, OpenACC, MPI, LLVM, …
- Libraries: BLAS, FFT, DNN, …