The HBP PCP: Technical Goals

The Human Brain Project

Dirk Pleiter | 18 December 2013

Purpose of this talk

- To outline our current thinking on what we will call for in the PCP tender
- To explain some of our thinking behind these ideas
- To get feedback from you as potential bidders

Today is a key opportunity for discussing these ideas prior to the formal PCP process

Outline

- HBP’s vision on Interactive Supercomputing
- Technical goals guiding R&D
- Pilot system to demonstrate readiness of developed technology
- Outlook on pre-exascale system

A vision of Interactive Supercomputing

- Future large-scale simulations within the HBP running on next generation, pre-exascale HPC platforms will require the possibility for data-intensive interactive simulations, analysis and visualization
- Spiking neuronal network simulations examples:
  - Interactive experiment design
  - Interactive data selection
  - Virtual surgery

Available roadmaps do not indicate appropriate solutions being developed

Envisioned mode of operation

- Users running multiple jobs concurrently within a single session
  - Dynamic change of job composition
- Job examples
  - Large-scale simulation jobs
  - Data analysis jobs
  - Visualization pipelines
- Job features
  - Multiple MPI processes
  - Multi-threaded processes
    (POSIX threads, OpenMP, OmpSs or other)
- Dynamic change of session resources

PCP targets only selected aspects of this vision

Key technical aspects

- Tightly integrated visualization
  - Rendering close to data, scalable image compositing
- Efficient data management
  - Significantly increased memory capacity to keep data within system
- Dynamic resource management
  - Dynamic relocation of resources within session and dynamic resizing of session resources
  - Co-scheduling of heterogeneous resources

Technical goals on dense memory

- Significant enlargement of memory capacity based on power-efficient memory technologies
  - Use cases
    - Enlargement of memory capacity available to the application
    - Buffer for transient data
    - Check-pointing in context of dynamic resource management
  - Target:
    Dense memory capacity / main memory capacity = 5
  - Suppliers are free in their choice of suitable memory technologies
    - Consider, e.g., hybrid approach using different memory technologies

Dense memory (cont.)

- Global, random addressable storage resource
  - Integration into network and accessibility from any node
  - Accessible via an RDMA interface
- Integration into a scalable network providing high bandwidth to near and far memory
  - Allow for check-pointing to near memory within O(1) minutes
  - Bandwidth to remote devices at most 10x slower under worst circumstances
- Management of dense memory as individual system resource

Technical goals on visualization

- Scalable rendering capabilities tightly integrated into a scalable HPC architecture with minimum data movement and copies
  - In-situ visualization solutions preferred
    - Data transfer within system to be minimized
    - Suitable benchmarks will be provided
  - Support of OpenGL 4.4 or later
- Scalable image compositing

Goals on resource management

- MPI level malleability
  - Temporal change of the number of nodes and/or processes used by an MPI application
  - Protocol to communicate with resource manager
- Job pre-emption
  - Classification of jobs wrt pre-emption capabilities
    - Interruptible, stoppable and non-interruptible
  - Migration of data to release main memory resources
  - Allow for modified number of processes after restart

Dynamic resource manager

- Expected functionality
  - Support of MPI level malleability
  - Handling of job pre-emption
  - Management of dense memory resources and visualization capabilities
- Technology to be developed on basis of current job management systems
  - Examples: SLURM, LSF, …

Technical goals on pilot system

- We expect real, working systems to be installed
  - Allow for demonstration of technology readiness
- System performance and usability verification
  - Application level benchmarks (NEURON, NEST)
  - Synthetic work-loads mimicking dynamic resource utilization and benchmarking viz sub-system
- Integration of dense memory technologies
- Scalable visualization sub-system
  - Depth based sort-last image compositing should be scalable to 4K UHD resolution at 30 fps for the full visualization subsystem size

Pilot system (cont.)

- Performance in terms of memory bandwidth
  - Plan for an aggregate $B_{mem} \geq 50$ TByte/s
- Large but balanced main memory capacity
  - Need large memory capacity C_{mem} to enable simulation of large spiking neuronal networks
  - Need high memory bandwidth B_{mem} to keep ratio of simulation time vs. simulated time small
  - Metric: memory capacity vs. bandwidth ratio $R_{mem} = C_{mem} / B_{mem}$
    - We consider to prefer $0.25s \leq R_{mem} \leq 0.75s$

Outlook on pre-exascale system

- We want confidence that technologies developed within PCP will become available as products
- Concept formulation with new technologies integrated
  - Integration of dense memory at scale
  - Tight integration of visualization
  - Scalable resource management
- Scalable architecture with up to 50 PFlop/s
- Supercomputer ready for use as discovery machine for neuroscience

Questions & Answers

- Before the PCP process starts we want to encourage dialogue and discuss

Please take the opportunity to ask questions and give feedback

- For questions and answers
  - During the following Q&A session(s)
  - Per email: hbp-pcp-dialogue@fz-juelich.de