Addressing the I/O bottleneck

Dr Michèle Weiland

NEXTGenIO Project Manager

EPCC, The University of Edinburgh

I/O is ***key*** Exascale challenge

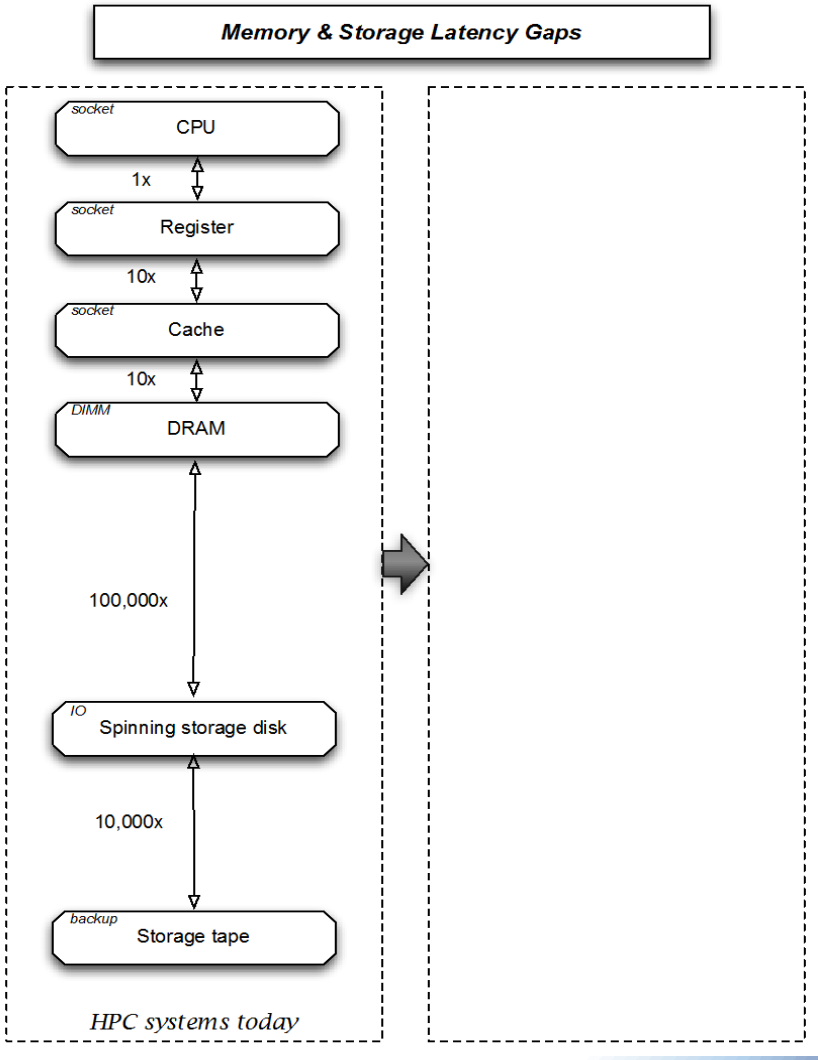
- Parallelism beyond 100 million threads demands a new approach to I/O
- Today's Petascale systems struggle with I/O
  - Inter-processor communication limits performance
  - Reading and writing data to parallel filesystems is a major bottleneck
- New technologies are needed
  - To improve inter-processor communication
  - To help us rethink data management and processing on capability systems

The “well balanced” computer

- Any computer system’s performance is limited by its slowest component
- For example
  - Reading from disk is often the slowest operation
  - We can add more disks in parallel until the aggregate disk throughput just saturates the CPU
  - … but this isn’t how many modern systems are designed with on-node disks rare in large systems
- *Amdahl number*
  - One bit of sequential I/O per second per instruction per second
  - Well-balanced ≅ 1, many HPC system today ≅ 10^{-5}

A new hierarchy

- 3D XPoint™ technology will profoundly change memory & storage hierarchies by bridging the latency gap
- HPC systems and Data Intensive systems will merge - HPDA
- Need to develop software – from the OS all the way to the application – to support NVRAM



NEXTGenIO summary

Project

- Research & Innovation Action
- 36 month duration
- €8.1 million
- Approx. 50% committed to hardware development
- Prototype system available from Month 27

Partners

- EPCC
- INTEL
- FUJITSU
- BSC
- TUD
- ALLINEA
- ECMWF
- ARCTUR

NEXTGenIO objectives

- Develop a new server architecture using next generation processor and memory advances
  - Based on Intel® Xeon and Intel DIMM based on 3D XPoint™ memory technology
- Investigate the best ways of utilising these technologies in HPC
  - Develop the systemware to support their use at the Exascale
- Model three different I/O workloads and use this understanding in a co-design process
  - Representative of real HPC centre workloads

Co-design process

- Hardware designers & integrators
- Technology providers
- HPC centres
- Tools developers
- Systemware developers
- Users and applications

Co-design process

- Architecture has 3 components:
  - Hardware
  - Systemware
  - Data

→ Key requirements: applications must be able to exploit NEXTGenIO platform without changes!

Applications

- Focus mainly on workloads, rather than specific applications
- Evaluation however will target specific applications and domains for verification and validation purposes
  - Weather and climate (IFS, MONC)
  - Engineering (OpenFOAM)
  - Visualisation (OPSray)
  - Chemistry (CASTEP)
  - Biological sciences (Roslin)

How can we exploit NVRAM?

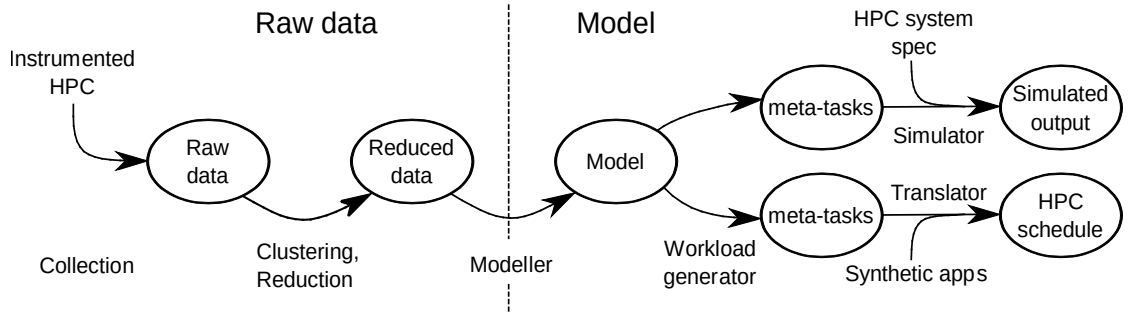
- Different use models
  - Check pointing of applications
    - Resiliency
    - Power efficiency
  - High performance parallel data storage
    - During job execution
    - Within a workflow
  - Very large memory applications
  - Intermediary storage

Job scheduler

- SLURM
  - Open source, proven performance
- Exploit NVRAM to optimise job flow through system
  - Pre-load data to where job will run
  - Write to disk after job has complete
- Enable booting of nodes into specific mode of operation for each job
  - Choose node configuration at job submission time

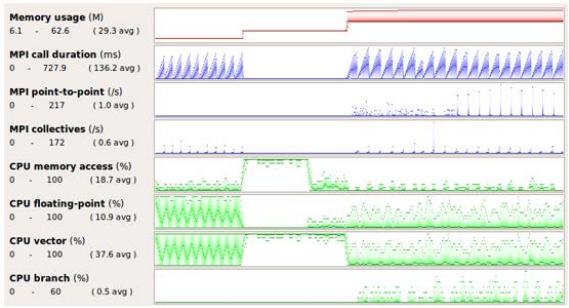
IO workload simulation

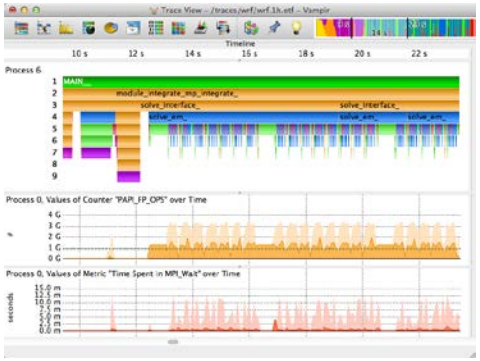
- Need to quantify improvements in job runtime and throughput
  - Measure and understand current bottlenecks
- Create a workload simulator and generator
  - Simulator can be used to derive system configuration options
  - Generator can be used to create scaled down version of data centre workload



Tools co-design

- Performance analysis tools need to understand new memory hierarchy and its impact on applications
  - TUD's Vampir & Allinea's MAP
- At the same time, tools themselves can exploit NVRAM to rapidly store sampling/tracing data





Final words

- NEXTGenIO is developing a ***full*** hardware and software solution
  - Real impact on future I/O
- Good progress, requirements capture and first architectural designs completed
  - Hardware under development
- Very exciting mix of hardware and software development
  - Co-design process extremely valuable to all parties