Addressing the I/O bottleneck of HPC workloads

Professor Mark Parsons

NEXTGenIO Project Chairman

Director, EPCC

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

Amdahl and 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 tried to quantify the characteristics of a well balanced computer in three “laws”

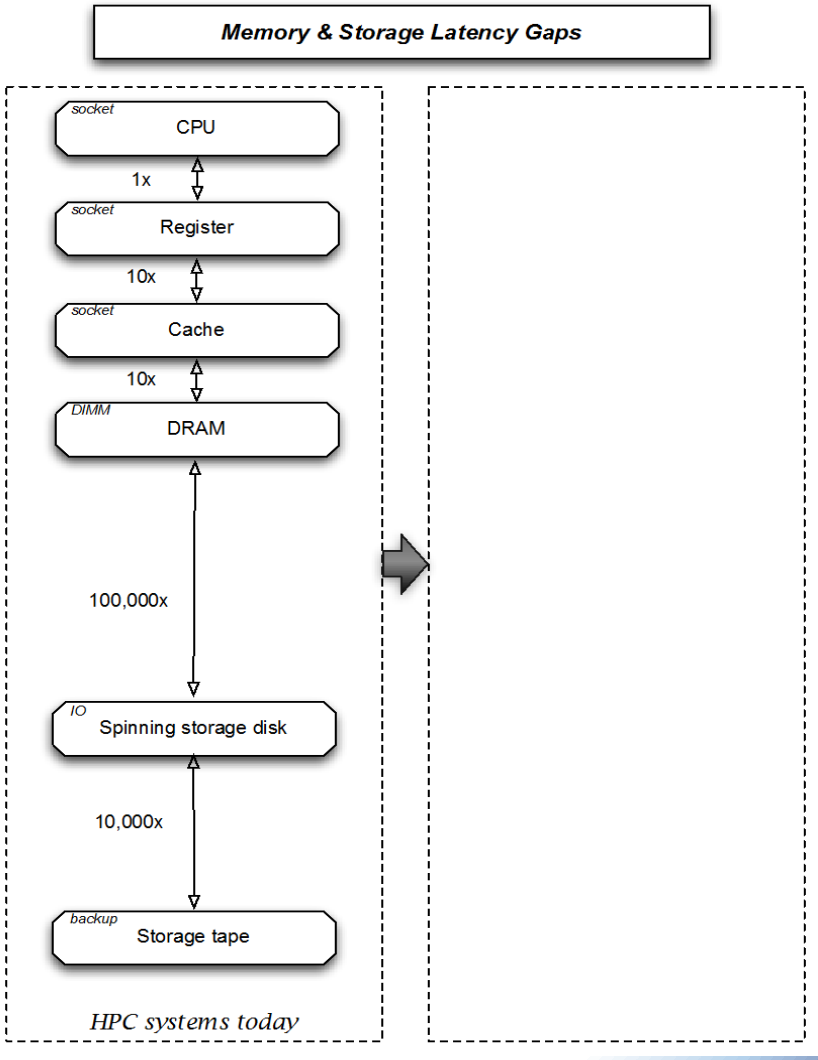
Three laws of a well balanced computer

Amdahl himself called these 'observations'

- Law 1
  - One bit of sequential I/O per second per instruction per second
  - This is called the *Amdahl number*
- Law 2
  - Has a memory with a Mbyte / MIPS ratio close to 1
  - This is called the *Amdahl memory ratio*
- Law 3
  - Performs one I/O operation per 50,000 instructions
  - This is called the *Amdahl IOPS ratio*
- A well balanced system today has Laws 1 and 2 ≈ 1
- Today for most hard disk technology Law 3 ≈ 0.014
- Many HPC systems have Amdahl numbers ≈ 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 non volatile RAM



NEXTGenIO summary

Project

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

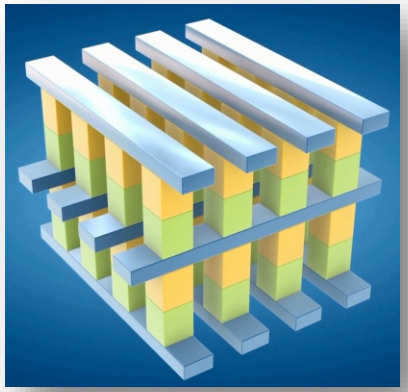
Partners

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

Intel DIMM using 3D XPoint™ Memory

Features

- Transistorless
- Very fast compared to NAND flash
- Low power (no DRAM refresh)
- Non-volatile
- Very large
- … and close to the CPU



NEXTGenIO objectives

- Develop new server architecture based on next gen Intel Xeon and 3D XPoint technologies
- Investigate how best to use it in HPC – develop the software stack
- Develop tool to model different I/O 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!

Real co-design?

- Some people say real co-design isn't possible – "it's a nice academic idea"
- NEXTGenIO ***isn't*** co-designing 3D Xpoint™ or the Intel Xeon (and chipset) we will use
- But ***it is*** co-designing the server motherboard and overall HPC system
  - What type of processors do we need to support?
  - How do we use the PCIe lanes?
  - Do we support Infiniband, Omnipath, or both or others?
  - Accelerators? … etc etc
- We are driving this process by thinking about what features our applications **and** systemware need

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)

Exploiting 3D XPoint™ in HPC

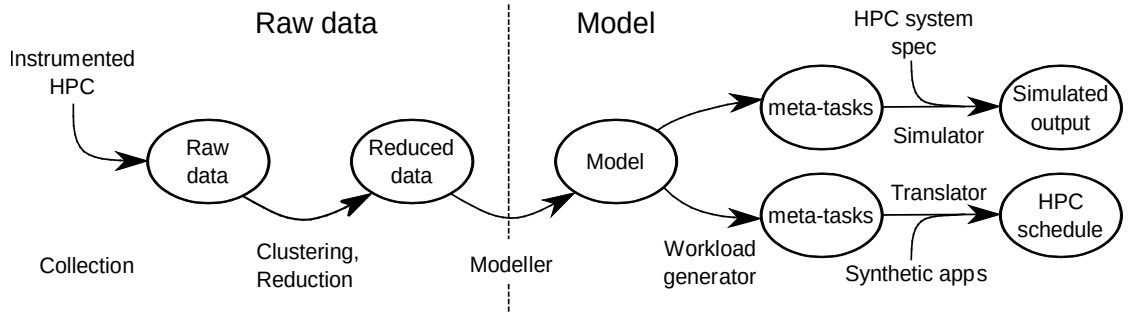
- Main options
  - As memory – volatile or non-volatile
  - As a file system
  - As a combination of the above
- Different use models
  - Check pointing of applications
    - Resiliency
    - Power efficiency
  - High performance parallel data storage
    - During job execution
    - Within a workflow
  - Very large memory applications
  - Intermediate 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

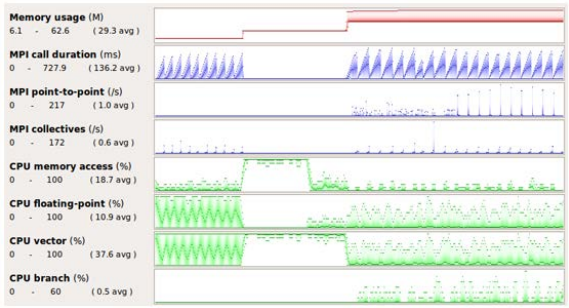
IO workload simulation

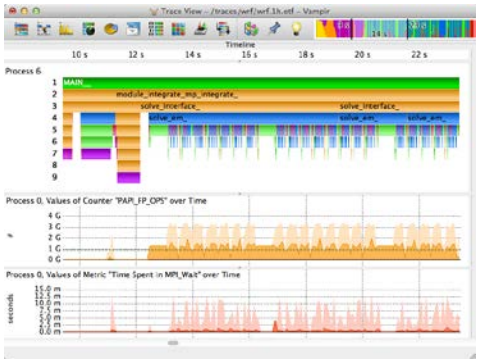
- Need to quantify improvements in job runtime and throughput
  - Measure and understand current bottlenecks
- Create a workload simulator and generator
  - Simulator can be used to derive system configuration options
  - Generator can be used to create scaled down version of data centre workload



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 HPC 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