The Exascale I/O Challenge

Professor Mark Parsons

Project Coordinator and EPCC Director

EPCC, The University of Edinburgh

Exascale is **very** challenging

- In 1990 EPCC's T800 based Meiko CS-1 delivered 800 Megaflops peak
- In 2015 our 118,080 core Cray XC30 delivers 2.5 Petaflops peak
- A 3.1 million times increase!
- Transition from Mega → Giga → Tera → Peta has been challenging but largely incremental
- We've lived in a golden age of stability
- But Exascale is much more challenging … 100+ million parallel threads …

Exascale is **very** challenging

- In 1990 EPCC's T800 based Meiko CS-1 delivered 800 Megaflops peak
- In 2015 our 118,080 core Cray … delivers 2.5 Petaflops peak
- A 3.1 million times in…
- Transition from … → Giga → Tera → Peta has been …ging but largely incremental
- We'… in a golden age of stability
- Bu… …xascale is much more challenging … 100+ million parallel threads …

I/O is a key Exascale challenge

Amdahl's Law

- S is speedup
- N is number of processors
- P is proportion of time code runs in parallel

$$S(N) = \frac{1}{(1-P) + \frac{P}{N}}$$

- For example:
  - If the code runs in parallel 90% of time then as N → ∞ the maximum speedup will be 10x

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 further laws

Three laws of a well balanced computer

- 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 $\approx$ 1
- Today for most hard disk technology Law 3 $\approx$ 0.014
- Many HPC systems have Amdahl numbers $\approx 10^{-5}$

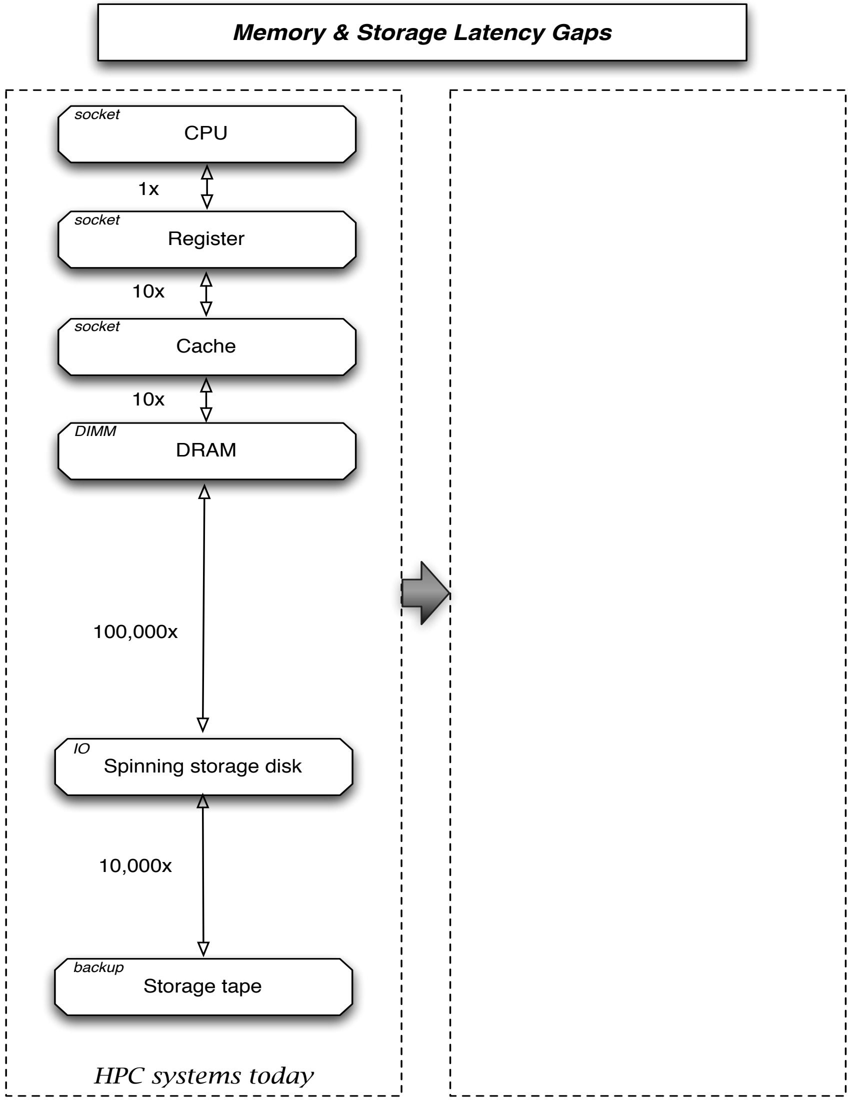
Three laws of a well balanced computer

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 $\approx$ 1
- Today for most hard disk technology Law 3 $\approx$ 0.014
- Many HPC systems have Amdahl numbers $\approx 10^{-5}$

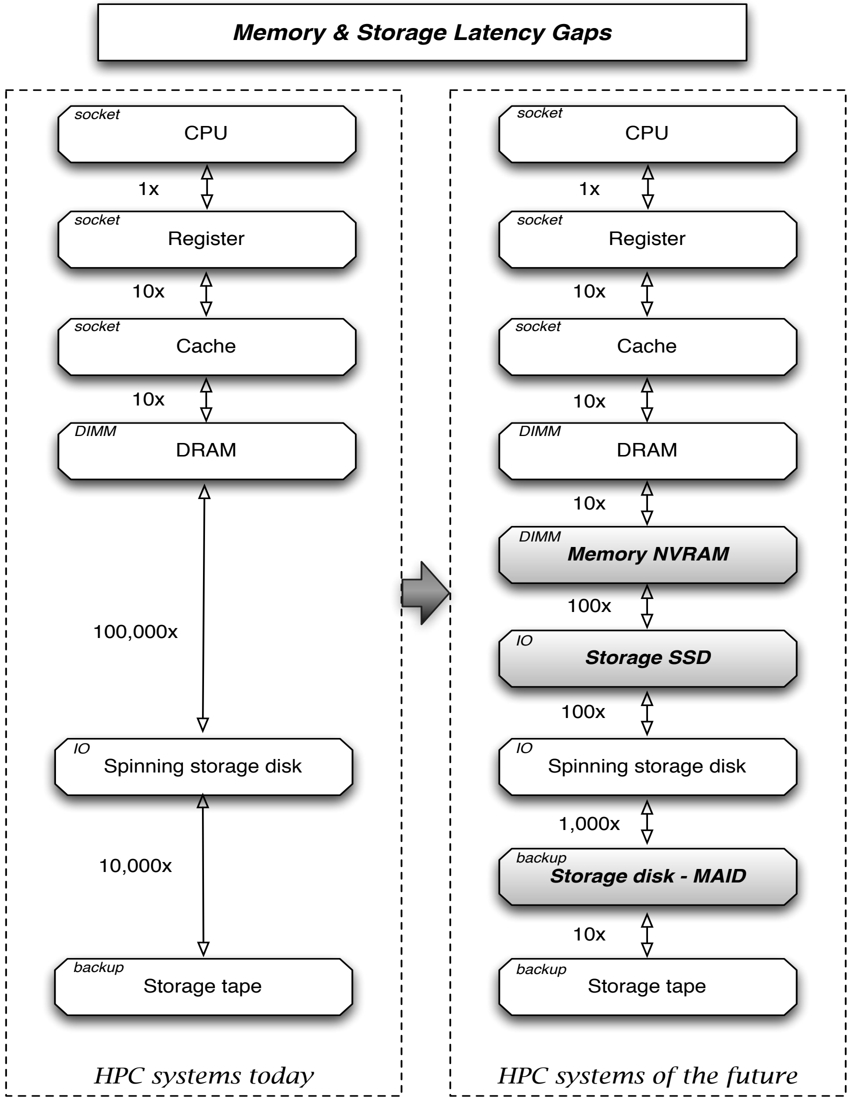
A new hierarchy

- Next generation NVRAM technologies will profoundly changing memory and storage hierarchies
- HPC systems and Data Intensive systems will merge - HPDA
- Profound changes are coming to ALL data centres
- … but in HPC we need to develop software – OS and application – to support their use



A new hierarchy

- Next generation NVRAM technologies will profoundly changing memory and storage hierarchies
- HPC systems and Data Intensive systems will merge - HPDA
- Profound changes are coming to ALL data centres
- … but in HPC we need to develop software – OS and application – to support their use



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

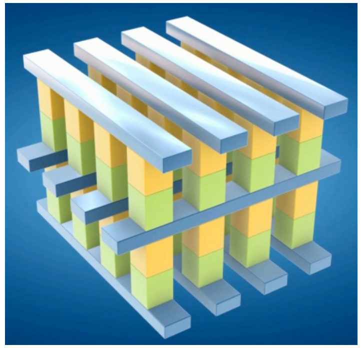
- EPCC
- INTEL
- FUJITSU
- BSC
- TUD
- ALLINEA
- ECMWF
- ARCTUR

I/O is *the* 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

NEXTGenIO objectives

- Develop a new server architecture using next generation processor and memory advances
  - Based on Intel Xeon and 3D XPoint technologies



- 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

Key Milestones

- M3 – Initial HW requirements available
- M6 – Initial HW architecture specification
- M7 – Tool selection and prototypes
- M15 – Power on of NV-DIMM samples
- M24 – Architecture finalised
- M27 – Hardware prototype delivered
- M30 – Systemware etc available on prototype
- M32 – Energy and data aware schedulers
- M36 – IO workload simulator released

How will we use this?

- 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

An example: ‘Hibernating’ an Exascale system

- A key Exascale challenge relates to electricity costs
- Early systems will require > 50Megawatts
- NV-DIMMs give us the opportunity to
  - ‘Barrier’ an entire system
  - Save all DRAM data to NV-DIMM
  - Power down during a peak period e.g. dinner time
  - Restart in a matter of seconds
- Easy to negotiate lower electricity pricing with this operational mode

Final words

- NEXTGenIO will be the first project to develop solutions using the 3D XPoint technology
- Very exciting mix of hardware and software development
- Strong team of partners
- Making good progress
- First architectural designs completed
- We agree this may be one of the most transformational projects any of us will ever work on