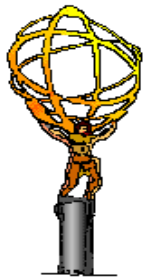
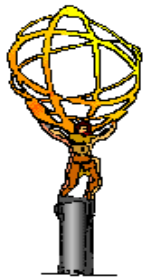


Experiment Requirements from Evolving Architectures

RWL Jones, Lancaster University

IC 9 July 2010



Changing Requirements

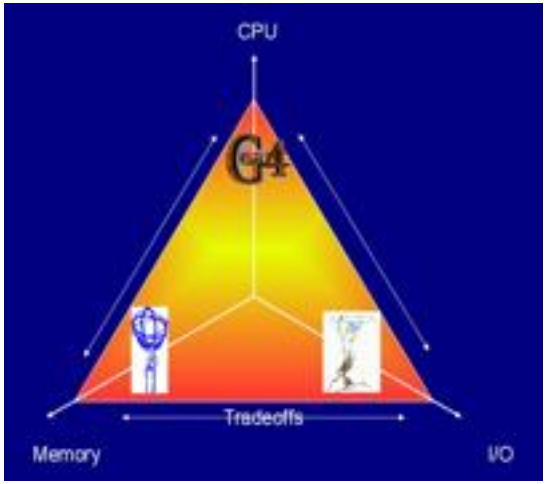
- Evolutions that are affecting our needs:
  - Multi core, many core
  - Virtualization
  - Clouds
  - GPGPU, APU……

Drivers:

- We are fighting
  - Increasing CPU need
  - Increasing memory demands per processor
  - Increasing bandwidth through processor
- Main increase in CPU power is from multi-core
  - This is leading to many core, which poses increasing challenges
- Possible alternate line – Graphical Processing Units
  - Not (yet) suitable for all applications
  - Things like offfline tracking may be among the ‘suitable cases for treatment’

Technology Challenge

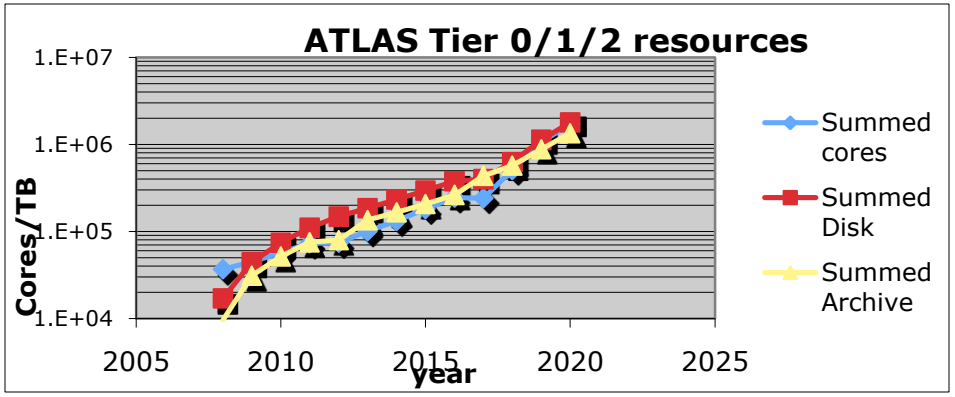
- Well known that the simulation and reconstruction of events for studies at 10^{35} (and even 10^{34}!) poses real challenges
  - Long time to process each event
  - High memory profile (presently more than total real memory on an 8-core processor)
- These all pose direct challenges to performance, which people are trying to address
- However, the challenges are much greater: Three components
  - Memory profile per process
  - I/O through processor
  - Data volume



- We are *obliged* to work smarter
- The Grid must support this

Data Future

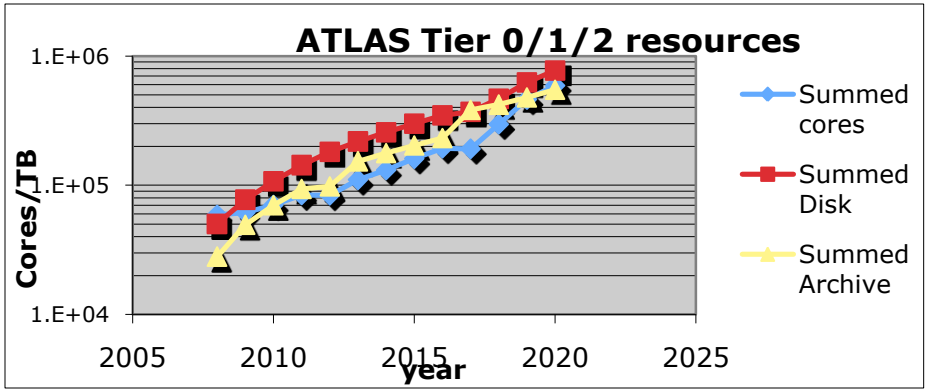
- Assuming trigger output flat rate until ~2014…



Scale output with specific lumi?

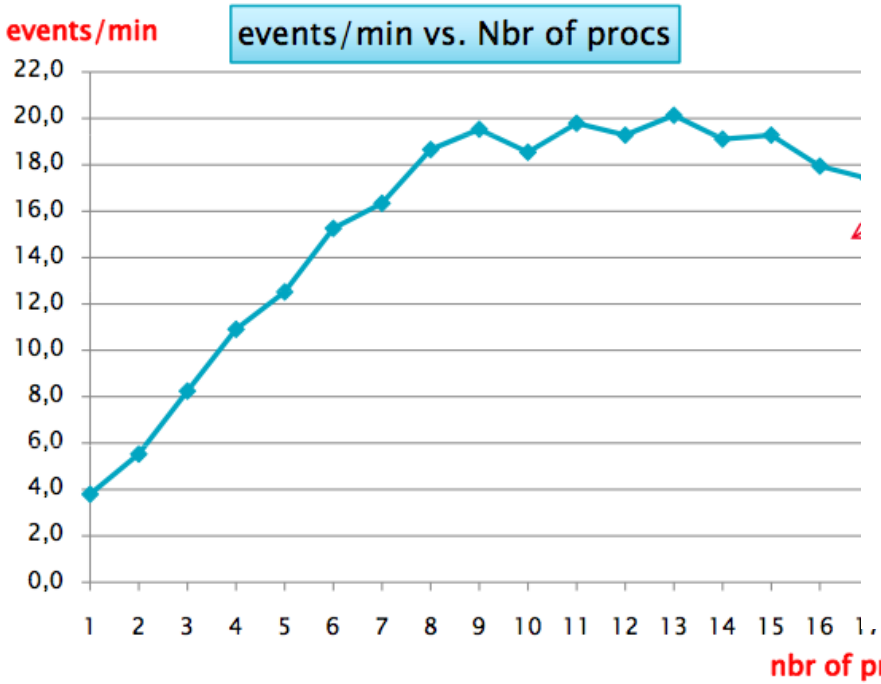
Fixed output for all lumi?

Follow a Moore's law growth
ASSUMING we can use the technologies



Parallelism
Initial Solutions: AthenaMP etc

- Generally work smarter! Event level parallelism
  - E.g. AthenaMP
    - Share common memory between parent and daughter processes to allow many on a single node
    - Some speed-up using event loop parallelism
  - Also share common pages between processes with KSM
    - Real gains in memory use, but some slow-down
  - Cache as much as you can (e.g. pile-up events)
  - Also Non-Uniform Memory Access, simultaneous multi-threading
- Issues: hard to monitor performance in parallel jobs

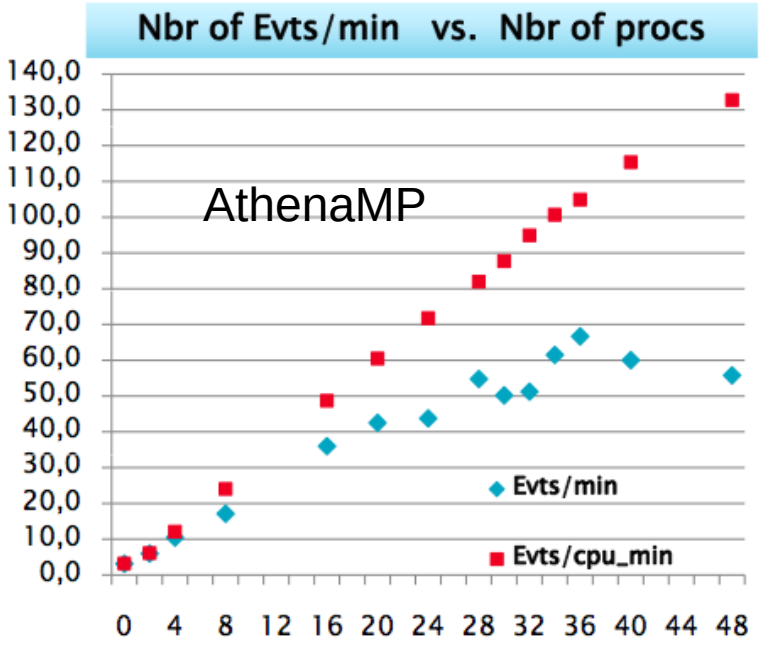


- Other approaches
  - Job level parallelism (e.g parallel Gaudi) & hyperthreading
  - Pinning of processes to cores or hyperthreads with Affinty

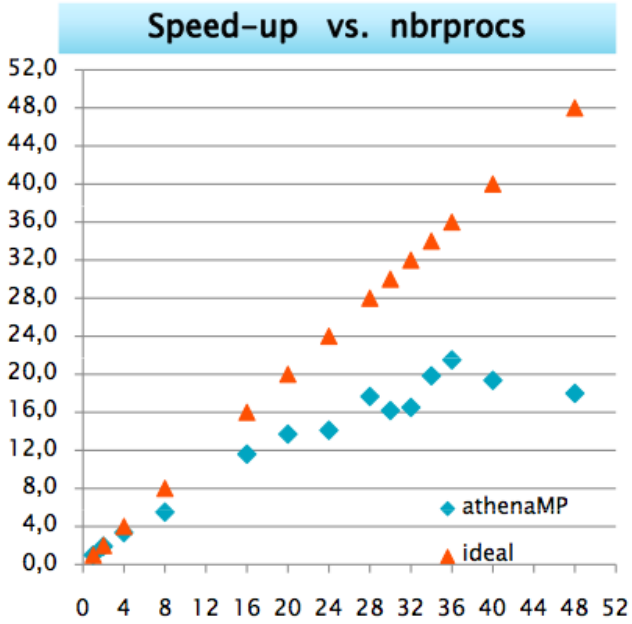
AthenaMP on 32 core NERSC machine

- Specs:
  - 8 x Quad-Core AMD Opteron™
  - Processor 8384 = 32 cores
  - 250GbofMemory! L2:512Kb,L3:6MB
  - Core Speed: 2700 MHz

Good (but)
Limits of approach evident
More parallelism needed



CPU scaling - speed-up



Related developments

- IO challenges being (partly) addressed by fast merging
- Re-write of Gaudi with stronger memory model planned
- Down the line, we need to parallise the code
  - This could be either for many-core processors or for Graphical Processing Units – but the development might address both
    - GPUs having big success & cost savings in other fields
    - Harder for us to use, but funders will continue to ask
    - We need the R&D to know which path to take
  - Developments require O(3 years) to implement
  - This includes Geant4 – architectural review this year

Multi-Core Workshops

- Second multi-core workshop 21/22 June
  - Under auspices of 2 OpenLab R&D projects
    - WP8
      Parallelization of Software Frameworks to explo it Multi-core Processors
    - WP9 Portable Analysis Environment using Virtu alization Technology
  - Experiments represented and gave a view of requirements in these two areas
  - Not complete, but important!

ATLAS Multi-core/Virtualization Requirements

- Need for a way to either schedule jobs taking over a complete computing node, or at least know how many slots out of the computing node one can take over. *Really* want the whole node
- The current VOBox service level should be based on VMs.
- A properly-working, properly supported CernVM and CVMFS would be valuable for ATLAS distributed analysis.
- Virtualization is no replacement for proper software distribution and configuration management, but it can make these tasks much easier and faster.

CMS Multi-core/Virtualization Requirements

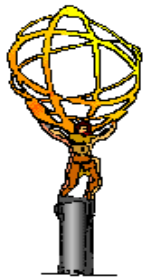
- CMS would like to proceed with commissioning/deployment of ” multi-core
  aware” applications in the next 6 months (by end 2010).
  - They propose moving to “whole node” scheduling as part of this commissioning.
- Desire storage systems & FTS to permit larger files, resolving problems with from-scratch restarts after errors.
- Wish to move beyond VSIZE/RSS for memory accounting
  - (PSS proposal [http://elmer.web.cern.ch/elmer/memory.html] “whole node” accounting needed)
- Virtualization appropriate for (most) services, but not as a per manent aspect of high-throughput WN's

LHCb Multi-core/Virtualization Requirements

- Policies must be put in place to use the parallel Gaudi and Athena on the Grid;
  - Multi-cored laptops and desktops can already .
  - Support is needed from the batch systems.
- CernVM should become a standard service supported by the b atch systems and transparent for the users.
- A solid virtualization infrastructure is required to be usable and the cost in terms of performance must be taken into account .

ALICE Virtualization Requirements/Comments

- Excellent experience with VMs for services for compacting rack space and a good environment for building, testing and proto typing.
- Multitude of adopted virtualization platforms (with their positive and negative sides)
- Mastering storage from a VM is still an open issue, especially data servers
- Virtualization is generally accepted for services that are not I/O demanding, also not for DBs.

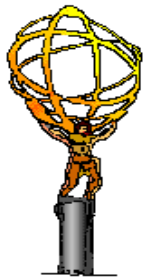


Summarizing Multi-core

- Experiments need nodes, not just cores
  - Experiments responsible for utilization, using
    - Pilot job mechanisms
    - Multiple processes and/or multiple threads
    - Job mix scheduling
- End-to-end changes needed in Grid frameworks
  - From user submission mechanism
  - To local batch system
- Accounting needs to change
- Larger files will result from parallel jobs

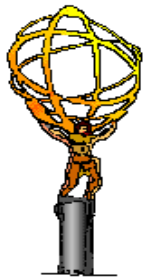
Summarizing Virtualization

- Virtualization of services is accepted and happens
  - VO boxes will be virtualized shortly
- Virtualization of WNs is for flexibility and efficiency
  - Performance hit must be known and monitored (e.g. IO to disk)
  - Cluster vitualization must support more than one image (e.g. Proof cluster, production,…)
  - HEPiX document to specify the obligations on image authors; separation of base OS and experiment code
  - Experiment support for CERNVM File System support for adding to image after instantiation
  - Some call for 24*7 CERNVM support
- Spin-out 3 discussion groups for next steps:
  - Performance, end-2-end roadmap, accounting



Personal questions

- Are the following cases being considered?
  - Virtualization for storage optimization
  - Virtualization on commercial clouds
  - Virtualization for live host migration

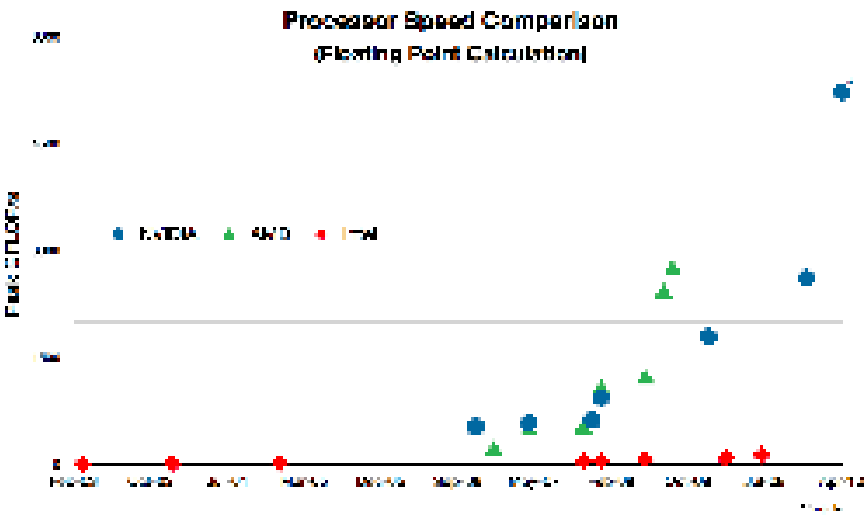


Other Issues: Storage developments

- Interaction with storage also challenging
  - Copy on write being investigated
  - Solid State Devices look promising for database access (TAG DB etc) and possibly analysis worker nodes
  - Need to change DB technologies for scaling
  - *Grid impacts minimal?*
- IO capacity does not scale with cores
- Facing issues in high-throughput access to storage (e.g. for analysis)
  - Common issue with HPC community
  - Emerging solution: data localization
  - Copy on write based technologies
  - *Needs changes to SE, CE and data management systems?*

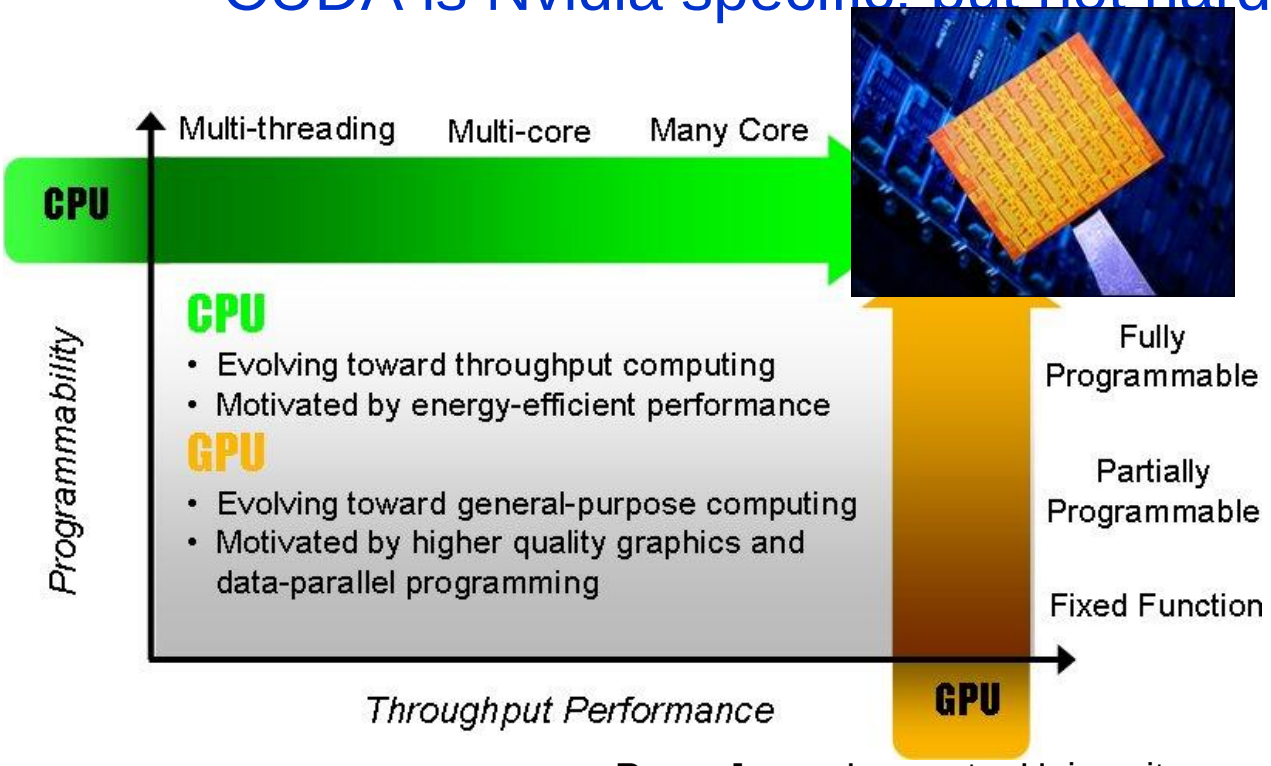
CPU vrs GPU vrs APU

- GPUs: 100's of cores, capable of simultaneously running O(10000) threads
- TFlops on a single GPU.
  - CPUs are optimized for low latency...
    - GPUs are optimized for high throughput.
- Not just faster than CPUs
  - Exponential increase in FLOPs
  - Lower power/cost
- Increasingly like CPUs
  - E.g. more double prec
- Must explore these
  - If these work for us, use them
  - If not, have an answer for FAs



Convergence

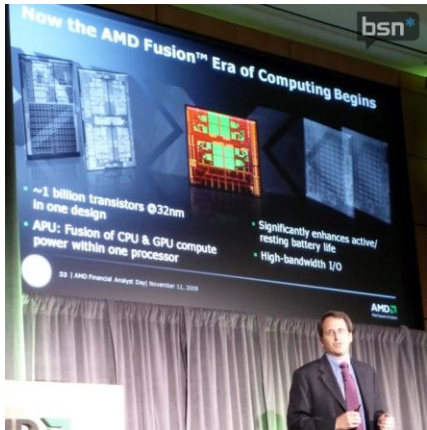
- Intel, AMD see a convergent future (APU)
- OpenCL provides development platform for both – avoid technology lock-in
- CUDA is Nvidia specific, but not hard to port to Open CL



Roger Jones, Lancaster University

48 core Single Chip Cloud unveiled 3/12/09

AMD Fusion unveiled To market 2011



Approach

- Most suitable problems
  - *Compute Intensity – Large number of arithmetic operations per IO or global memory reference.*
  - *Data Parallelism –same function applied to all records of an input stream and no need to wait for previous records.*
  - *Data Locality –data produced , read once or twice later in the application, and never read again.*
- ATLAS specific demonstrators
  - Tracking code (examples exist from other experiments)
  - Magnetic field service (could save memory and well suited to GPU service)
    - Run service in GPU as co-processor
- CEs must support
  - GPUs as primary compute node – many issues same as multicore
  - GPUs as co-processor
  - Work starting in UK on this, level of effort uncertain!

Conclusion

- Offline now has Upgrade Computing activity area
  - Focus of development, LoI etc
  - Suitable forum for trigger/offline constraint planning?
- A-team meetings for regular technical exchange
- Is this enough?
  - Suggest review later this year after initial experience