Future Computing

Graeme Stewart, Roger Jones

University of Glasgow | Department of Physics & Astronomy

LANCASTER UNIVERSITY

Clouds

- Now many cloud services online
- Flexible compute and storage
  - If you can pay
- Probably useful for simulation
  - ~US$1 for 50 events on EC2/S3
    - 100M events costs $2M
      - 2000 slot T2 can do ~100M events in a year
    - Data costs for reco and analysis prohibitive
      - 500TB in S3 is US$30k/month
      - 500TB of data movement is US$40k
        - 1 year with 4 refreshes costs US$520k
- Maybe an offer from CERN-IT to help pay the bill?
- Treat cloud as panda site
  - Assign a G4 task and let it run

Clouds II

- Practical issues:
  - CE/batch system in the cloud (e.g., condor EC2) or pilot built in to VM? CVM co-pilot work can be used
    - Security issues
    - Data staging: from out of cloud or using cloud storage
      - The later more ambitious but probably cheaper
      - Stage-out
      - S3 as DDM site
        - Require support from FTS?
- Not just Amazon and friends:
  - Projects in UK (Edinburgh), US (Argonne), etc.
  - Common VM should be made (CERN VM, built-in pilot?)

Multi-Cores/Highly Parallel

- Already 6 core Intel CPUs common, 12 code AMD Magny Cours.
  - 48 cores in a single chassis
  - Maybe 50+ cores by end of 2011
- Experiment requests for whole node scheduling: test queues at CERN, SARA, RAL
  - Memory sharing
  - Coherent i/o
    - New storage models needed?
    - New i/o framework to support?
  - Larger outputs (e.g., 500 events and 400MB HITS from G4)
- Athena 16.5.0.1 supports multi-threaded reconstruction
- Need to test workflows now
  - We seem to have many pieces, but need to put them together

Multi-Cores/Highly Parallel

- Later on:
  - Workload sharing, with CPU intensive backfill
- Graphics processors/coprocessors?
  - Unlikely on most sites in 3 years
  - But GPU farms for specialist analysis tasks (~1 per cloud) plausible within 3 years
  - Root or Athena? Grid access will be needed
  - Directives?
- Whole node meshes better with cloud IaaS

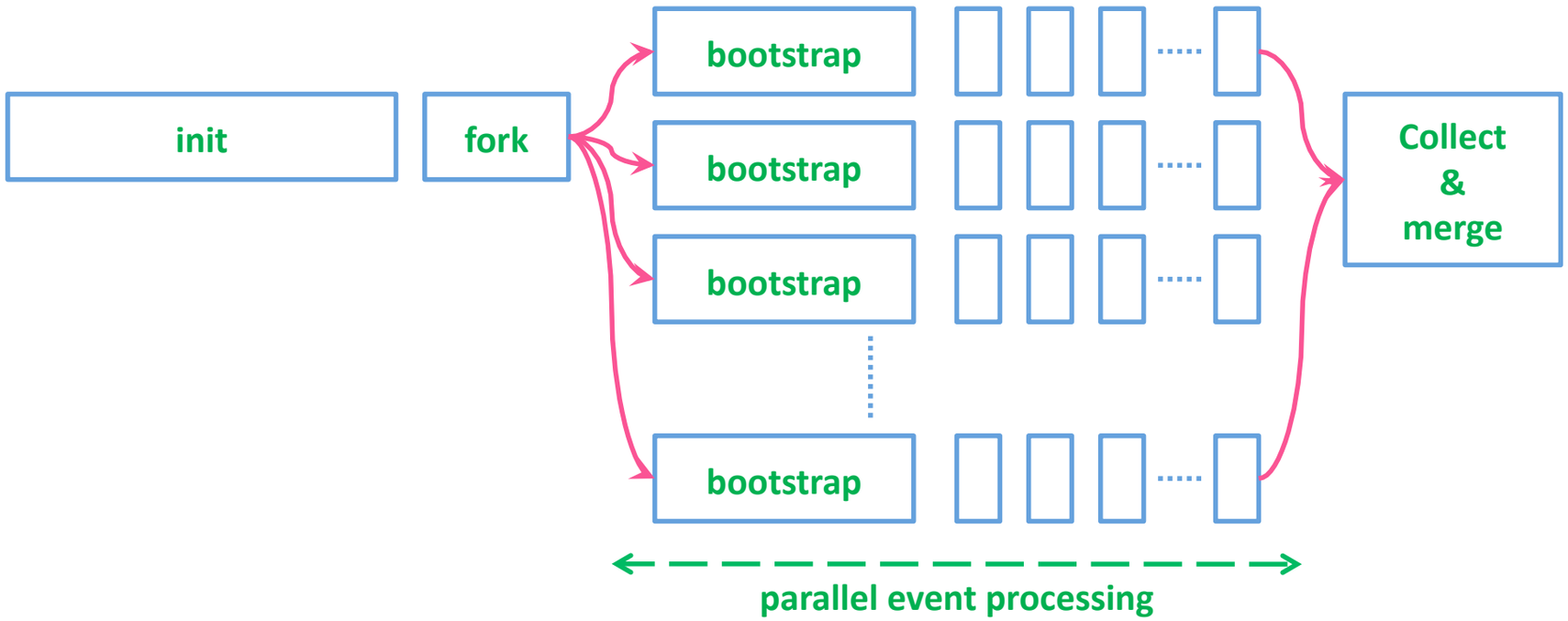
Compute Infrastructure at Sites

- Traditional job processing:
  - Batch system scheduling work
  - Grid layered on a CE: basically an X509 auth layer on top of a dumbed down qsub
    - Little conceptual change
- Pilot jobs evolved to
  - Circumvent extraneous layers and improve reliability
  - Provide global scheduling
- Late binding of jobs, plus whole node scheduling
  - Who needs a CE anymore?
    - Direct pilot submission to queues
    - VM load balancer for VMs with pilots built in
  - How to manage feedback with sites?
    - Pilot factory as a pressure sensor

Other Topics

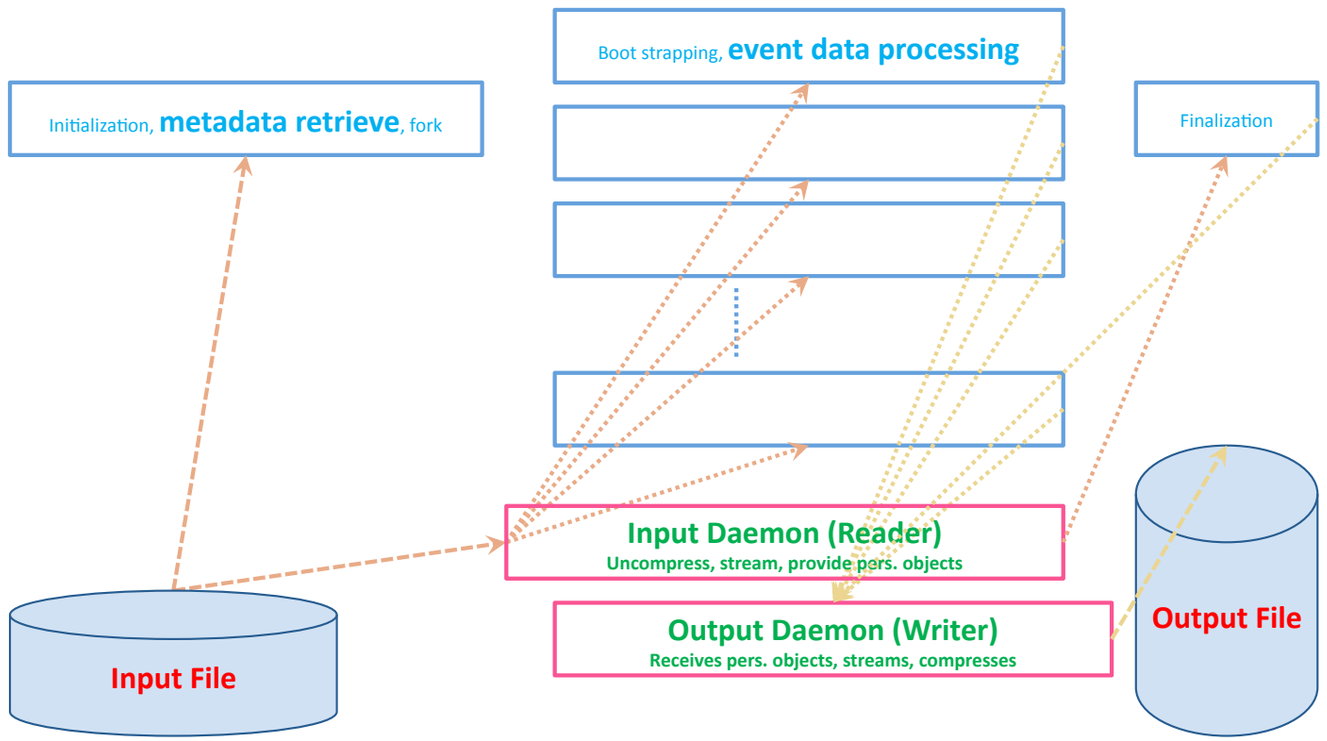
- MapReduce
  - Industry paradigm for massive data analysis
    - Use, e.g., HDFS storage
- Need data blocked on event boundaries
  - Convert root files to key-value store?
- Need lightweight startup for athena
  - Dynamic plugin for use code?

IO Frameworks



- Current IO framework is post-hoc
  - Each node produces output, later merged
  - Can now be done faster with POOL/ROOT fast merge outside Athena (but no metadata update, no updated references, non-optimal compression)

A better framework?



- Current Athena
  - All workers share same file on read
  - All copies of same baskets uncompressed separately
  - Compression done non-optimally, one per writer
- Proposed scatter/gather would address
  - But what to scatter when?
  - One scatter/gatherer per node? Per file?
- Other models are possible
  - Concurrency control? Resource sharing?