Computing Directions

Ian Fisk and Jim Shank

July 30, 2013

What follows

- Jim Shank and I are tasked with developing the Computing Report for the Energy Frontier
  - Computing doesn't drive the research program, but it does enable it.
  - Looking at the machine plans. They are all high luminosity machines with potentially very high trigger rates and complicated events
  - Constrained budgets, and few miracle solutions
- What follows are some observations to spawn discussion

Looking Back

- We decided to look back 10 years before trying to look forward 10
  - Tevatron was in the 3rd year of Run2 in 2003
- Compare to 2012
  - The third year of LHC
- What it shows you is that new machines can lead to big jumps in some resources

Complexity and Collaborations

- Trigger rate, event size, and reconstruction time all rise by a factor of 10

Metric	Tevatron (2003)	LHC (2012)
Trigger rate	50Hz	500Hz
Prompt Reconstruction rate/week	13M Events	120M events
Re-reconstruction rate	100M events per month	800M – 1B events per month
Reconstructed size	200kB	1-2MB
AOD size	20kB	200-300kB
Reconstruction time	1-2s on CPUs of the time	~10s on CPUs of the time

- Collaborations increase by a factor of 3

Metric	Tevatron(2003)	LHC(2012)
Collaboration Size	800	2000-3000
Number of individual analysis submitters per day	100	300-400
Number of total analysis submitters	400	Greater than 1000

Ian Fisk FNAL

Resources

- Resources and challenges increase at different rates

Metric	Tevatron(2003)	LHC(2012)
Remote Computing Capacity	15kHS06 (DZero Estimated)	450kHS06 (CMS)
User Jobs launched per day	10k per day	200-300k jobs per day
Disk Capacity per experiment in PB	0.5PB	60PB
Data on Tape per experiment	400TB	70PB
MC Processing Capacity per month for Full Simulation	3M	300M
Data Served from dCache at FNAL per day	25TB per day	10PB per day
Wide Area networking from host lab	200Mb/s	20000Mb/s
Inter VO transfer volume per day	6TB (DZero SAM)	546TB (ATLAS)

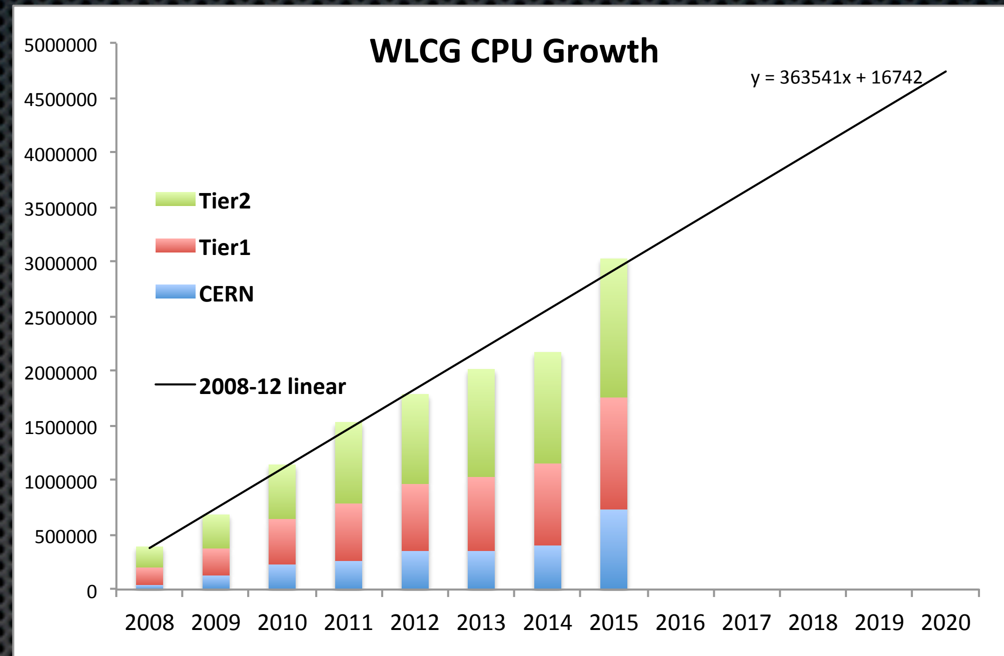
Ian Fisk FNAL

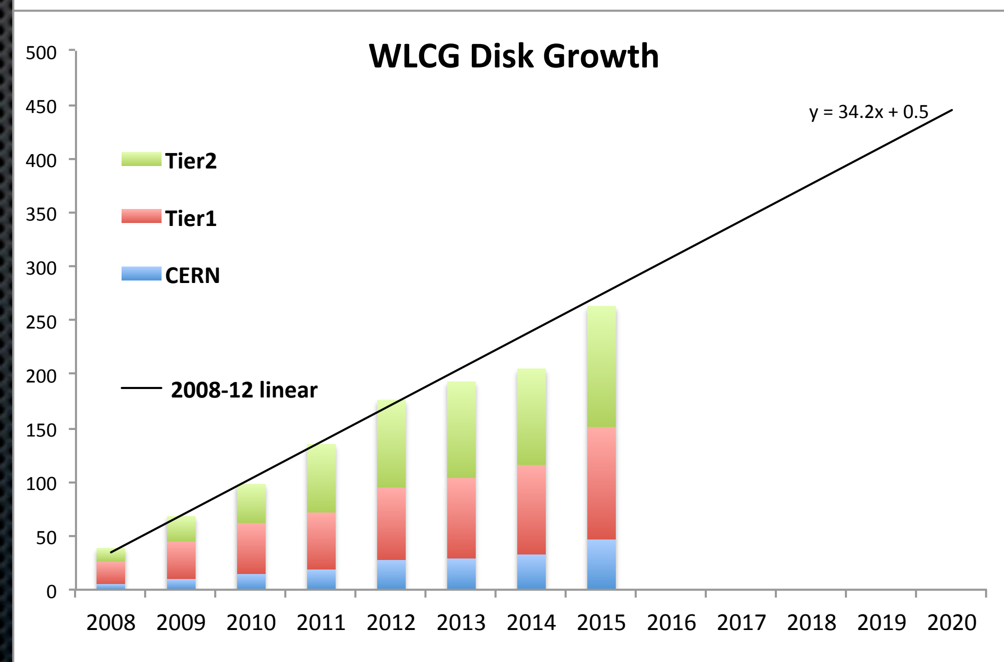
Increases

- The processing has increased by a factor of 30 in capacity
  - This is essentially what would be expected from a Moore's law increase with a 2 year cycle
  - Says we spent similar amounts
- Storage and networking have both increased by a factor of 100
  - 10 times trigger and 10 times event size

For LHC Increases per year

- LHC Computing adds about 25k processor cores a year
- And 34PB of disk
- The χ^2 of the linear fit is not very compelling, but it shows its currently increasing at a sustainable rate
    - A decade from now would be a factor of 4-5 in capacity





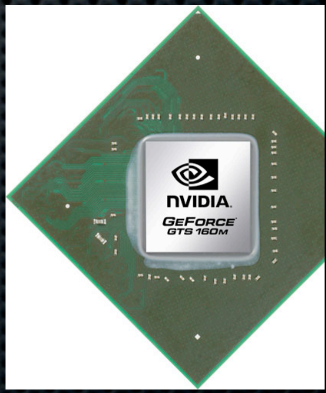
Ian Fisk FNAL

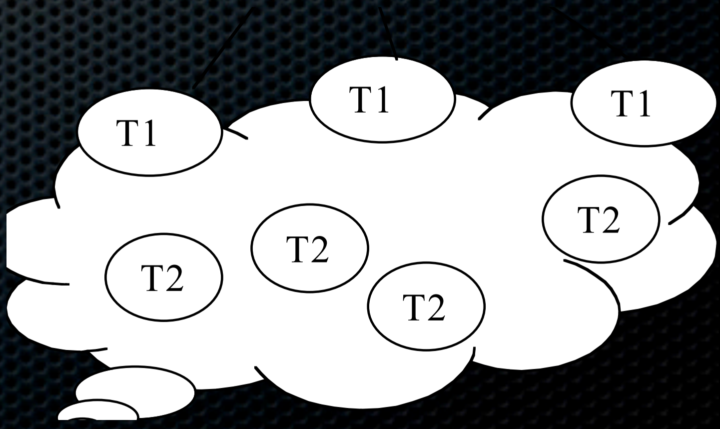
Looking Forward

- The programs suggested for energy frontier all have the potential for another factor of 10 in trigger and 10 in complexity
- Simulation and reconstruction might continue to scale with Moore's law as they did for LHC, but could just as easily increase much faster
- How to make better use of resources as the technology changes?

Looking Forward

- Computing is at something of a cross roads
  - In one direction are clouds
    - Generic computing services that are bought, shared, or contributed
    - Computing as a service
  - In the other direction are very specialized systems
    - High performance, low power
      - Massively multi-core
      - GPUs





Ian Fisk FNAL

Clouds and Provisioning

- Commercial clouds are still very expensive for resources we use a lot
  - Small sites without a history of computing will probably be the first to simply buy capacity
- More opportunistic and academic resources will move to cloud provisioning methods
- Even sites we control will move to cloud provisioning tools because it simplifies the operations and places more expectations on the supported community to define and operate services
- We should expect our current service architecture will change to new provisioning tools

Service Architecture

- We should expect our current service architecture for accessing resources will become a lot more diverse
  - In LHC Run 1 we enjoyed a lot of consistency with the WLCG
  - Looking forward we will have Cloud interfaces, Grid, local, opportunistic, and whatever comes next
  - We will need to have systems that do resource provisioning on all of it, and make it look like a coherent system

Current Hardware

- Currently energy frontier computing lives in a homogenous but non optimal environment
  - Looking back we have typically supported many more platforms
  - Most of the industry development is not in the chips that make up the bulk of our computing
  - Cores are added, but individual cores tend to stay at similar speed, with the exception of power efficiency there is not the same incentive to replace gear
  - We are not well optimized and we don’t tend to use the full capacity of the hardware

Ian Fisk FNAL

Specialized Hardware

- Specialized chips like GPUs and co-processors have the potential for big improvements in performance, but are challenging to program and introduce a lot of heterogeneity
- Specialized machines like very high core count low power systems look like super computers
  - And have the programming challenges associated

Specialized Hardware Steps

- 1.) We buy/get access to specialized gear like a super computer allocation
  - Big offline applications seems like the ideal use-case for buying or contributed capacity. Technology situation changes rapidly
- 2.) Places we completely control will get specialized gear for individual applications, but probably for niche applications
  - Trigger farms and other specialized use cases

Data Management

- We will have a mix of local, cloud, opportunistic, and specialized resources and we will need a data management system that deals with all
    - On cloud the concept of data locality begins to lose a lot of meaning
    - We cannot really afford another factor of 100 increase in storage, so we need to find ways of being more efficient in the use of the space
- We need to identify technology that allows a system to distribute and serve the data much more flexibly and dynamically

Connectivity

- Given the connectivity of our clusters and the expectations of the users, I believe we will have to evolve to content delivery networks
  - Data Management resources that deliver data on demand
  - Will be cached and replicated and intelligent about the placement, but large independent local storage systems connected to clusters is probably not the most efficient
- The data federations already being deployed are a first step, but work is needed

Networking

- Data delivery systems give a lot of flexibility in terms of how to make use of diverse computing systems, but they put strong requirements on networking
  - Currently a 10k core cluster (typical for 2020) would require 10Gb/s networking for organized processing like reconstruction
  - Analysis would require 100Gb/s
-

Becoming More Selective

- We have not really changed how we think about events we select
  - Currently we make a trigger decision and then all events are equal
  - Trigger rates continue to rise with intensity and most events are uninteresting background
- We may be able to afford to write things to tape, but may want to reduce the actively analyzed data

Not all events are equal

- A decision that is given 100ms of thought does not have to be the final word
  - We should be prepared to reduce our active dataset through reprocessing and understanding the data
    - Many things may be classified as known physics and put into distributions, but not kept in the active dataset
    - It's the equivalent of what is done in analysis, but in a more organized way
- We can afford a lot of data on tape, but the active dataset is much more expensive

Energy Frontier

- As trigger rates proposed for Energy Frontier approach rates we would typically associate with Intensity Frontier we may need to adopt similar techniques
  - ALICE is already planning for this post LS2.   Much more immediate processing and identification

Outlook

- LHC moving forward may be sustainable with an evolution of how we work
- A big increase in luminosity and complexity would lead to a big jump that would be potentially very expensive
    - To handle this we need to change how we work by being more selective
    - Move to be able to run on fast hardware
    - Solve the data management problem