Fermilab U.S. DEPARTMENT OF ENERGY Office of Science

High Performance Computing/Massively Parallel Computing Activities and Developments at FNAL: an Overview

James Amundson, *Fermilab*

8th INFIERI Workshop

October 20, 2016

Computing in High Energy Physics

HEP computing involves the processing of ever larger numbers of independent events.

"Trivially parallelizable"

I'll leave details for other talks, but:

- The top quark was discovered at Fermilab in 1995
  - roughly 1 in 1,000,000,000,000 Tevatron collisions produced a top quark
- The Higgs Boson was discovered at CERN in 2012
  - supporting evidence from Fermilab (Tevatron)
  - roughly 1 in 100,000,000,000,000 LHC collisions yielded a distinguishable Higgs Boson
- Plans for HL-LHC include a 100x increase in data

A little historical background before looking forward

HEP computing has gone through several different eras

- B.C.
- ...
- VAXus Vulgaris
- Unix Principium
- Linux Maximus

and, looking to the future,

- Deus Ex Machina

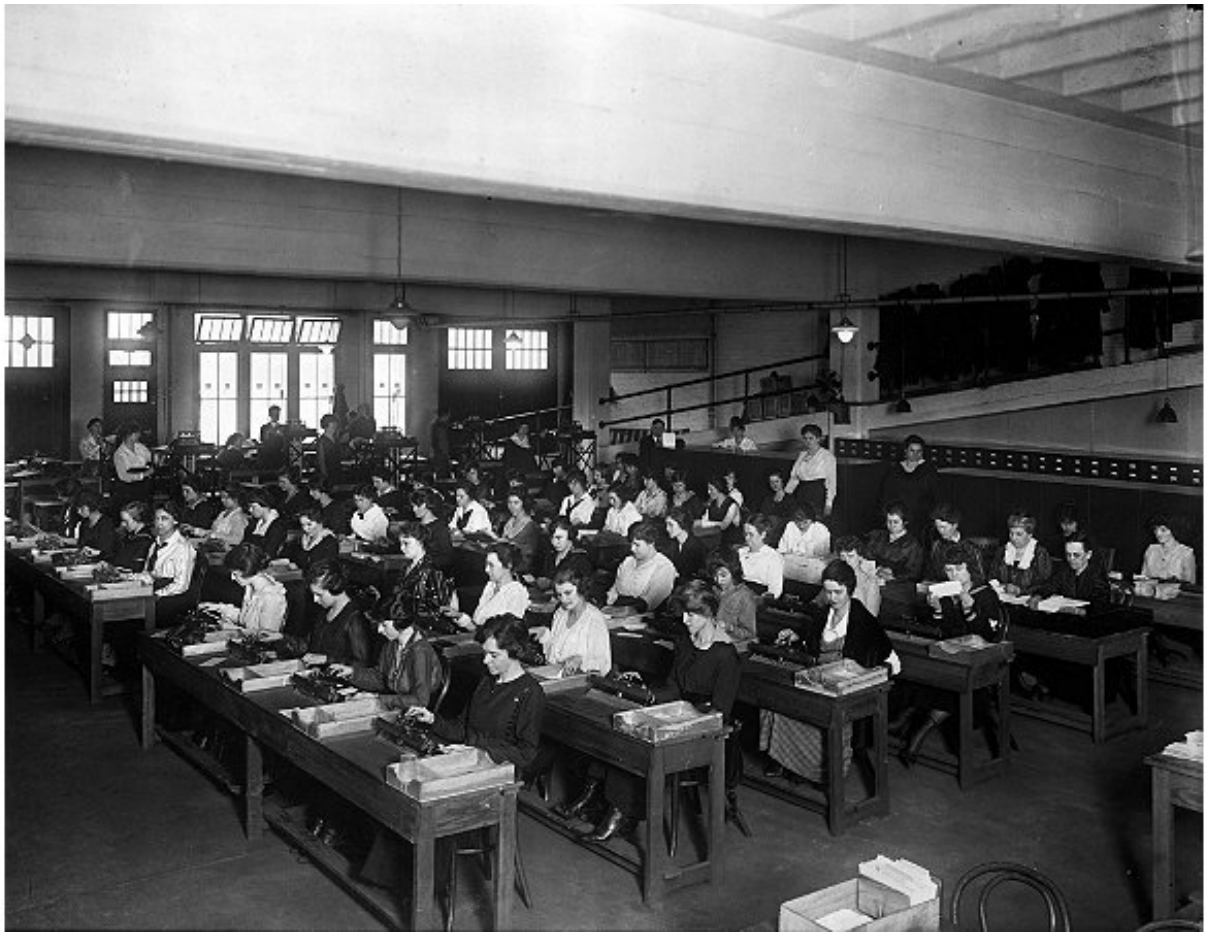
Era: *B.C.*

Before Computers

From an *Atlantic* article entitled

“Computing Power Used to Be Measured in 'Kilo-Girls’”

Presumably now we would use “Kilo-Grown Women.”

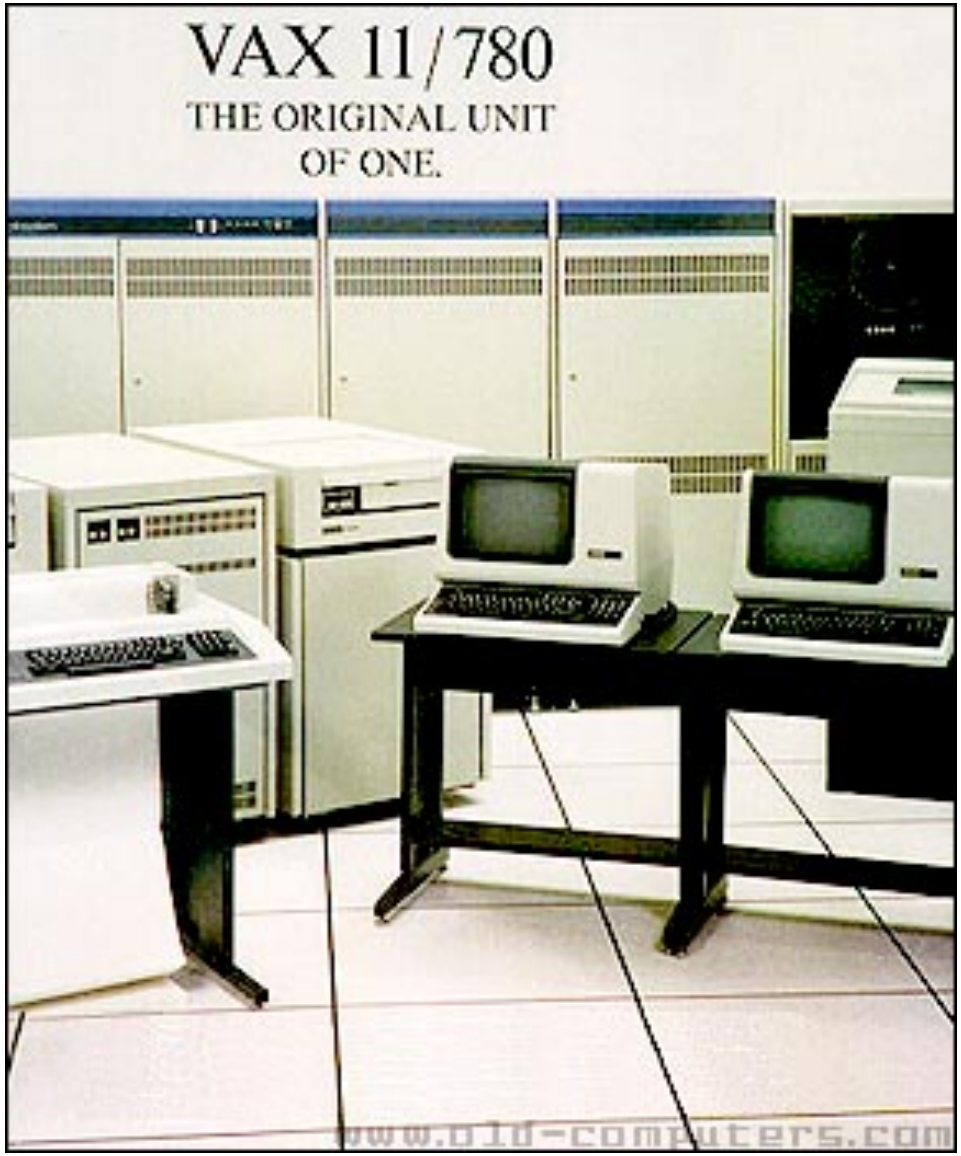


Women at work tabulating during World War II (Shorpy)

Era: *VAXus Vulgaris*

(skipping ahead to the 80's…)

DEC VAX with the VMS operating system was the most popular HEP computing platform in the 80's.



Era: *Unix Principium*

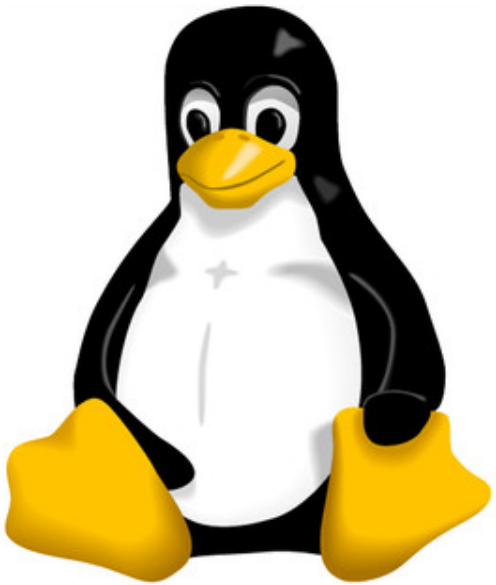
The 90's saw a transition to Unix-based computers

- Many people were skeptical that physicists could ever move from VMS to Unix.

- Nonetheless, Unix workstations and larger shared-memory machines came to dominate.

Current Era: *Linux Maximus*

Since the early 2000's, HEP has been dominated by large clusters of x86-based Linux machines.



- Many people were skeptical that physicists could work without large shared-memory machines.

- HEP was an early adopter of Linux clusters. Google, Amazon and others now have clusters that dwarf ours.

HEP computing in context

HEP requires High Throughput Computing (HTC).

- Large numbers of independent computations.

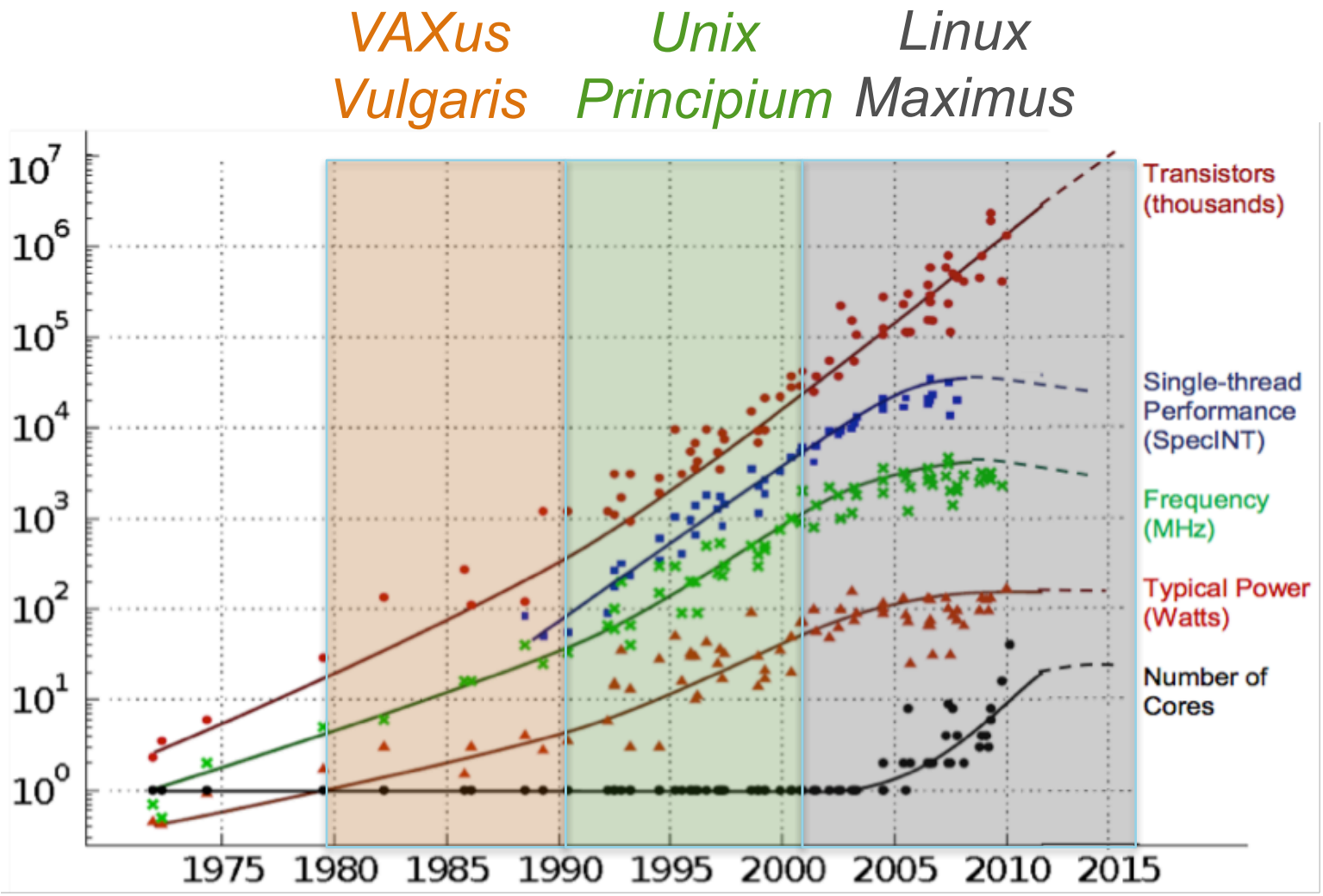
Many other areas of computational science require High Performance Computing (HPC).

- Large, tightly coupled calculations.
- Typically performed on supercomputers, or Linux clusters with specialized networking.

The world of HPC: top500.org

Rank	Site	System	Cores	Rmax (TFlop/s)	Rpeak (TFlop/s)	Power (kW)
1	National Supercomputing Center in Wuxi China	**Sunway TaihuLight** - Sunway MPP, Sunway SW26010 260C 1.45GHz, Sunway NRCPC	10,649,600	93,014.6	125,435.9	15,371
2	National Super Computer Center in Guangzhou China	**Tianhe-2 (MilkyWay-2)** - TH-IVB-FEP Cluster, Intel Xeon E5-2692 12C 2.200GHz, TH Express-2, Intel Xeon Phi 31S1P NUDT	3,120,000	33,862.7	54,902.4	17,808
3	DOE/SC/Oak Ridge National Laboratory United States	**Titan** - Cray XK7 , Opteron 6274 16C 2.200GHz, Cray Gemini interconnect, NVIDIA K20x Cray Inc.	560,640	17,590.0	27,112.5	8,209
4	DOE/NNSA/LLNL United States	**Sequoia** - BlueGene/Q, Power BQC 16C 1.60 GHz, Custom IBM	1,572,864	17,173.2	20,132.7	7,890
5	RIKEN Advanced Institute for Computational Science (AICS) Japan	K computer, SPARC64 VIIIfx 2.0GHz, Tofu interconnect Fujitsu	705,024	10,510.0	11,280.4	12,660
6	DOE/SC/Argonne National Laboratory United States	**Mira** - BlueGene/Q, Power BQC 16C 1.60GHz, Custom IBM	786,432	8,586.6	10,066.3	3,945

Trends in computing hardware



"Data Processing in Exascale-Class Computing Systems", Chuck Moore, AMD Corporate Fellow and CTO of Technology Group, presented at the 2011 Salishan Conference on High-speed Computing, Original data collected and plotted by M. Horowitz, F. Labonte, O. Shacham, K. Olukotun, L. Hammond, and C. Batten, dotted line extrapolations by C. Moore

Observations on hardware trends

- Single-thread performance increased steadily from the beginning of the Unix era through the beginning of the Linux era
  - Multi-core processors have since taken up the slack
  - HTC has required only minimal adaptation
  - Memory per thread has become stagnant

- Trends that have persisted for decades have ended (e.g., increases in clock speeds) and new ones are beginning (e.g., increases in core counts).

“May you live in interesting times”

Not logged in Talk Contributions Create account Log in

WIKIPEDIA
The Free Encyclopedia

Main page
Contents
Featured content
Current events
Random article
Donate to Wikipedia
Wikipedia store

Interaction
Help

Article Talk Read Edit View history Search

May you live in interesting times

From Wikipedia, the free encyclopedia

> *"Chinese curse" redirects here. For Chinese-language profanity, see Mandarin Chinese profanity.*

"**May you live in interesting times**" is an English expression purported to be a translation of a traditional Chinese curse. Despite being so common in English as to be known as "**the Chinese curse**", the saying is apocryphal, and no actual Chinese source has ever been produced. The most likely connection to Chinese culture may be deduced from analysis of the late-19th century speeches of Joseph Chamberlain, probably erroneously transmitted and revised through his son Austen Chamberlain.[1]

Future era: Deus Ex Machina

President Obama, July 29, 2015:

EXECUTIVE ORDER

CREATING A NATIONAL STRATEGIC COMPUTING INITIATIVE

By the authority vested in me as President by the Constitution and the laws of the United States of America, and to maximize benefits of high-performance computing (HPC) research, development, and deployment, it is hereby ordered as follows:

…

Sec. 2. Objectives. Executive departments, agencies, and offices (agencies) participating in the NSCI shall pursue five strategic objectives:

1. **Accelerating delivery of a capable exascale computing system that integrates hardware and software capability to deliver approximately 100 times the performance of current 10 petaflop systems across a range of applications representing government needs.**

…

Exascale computing challenges

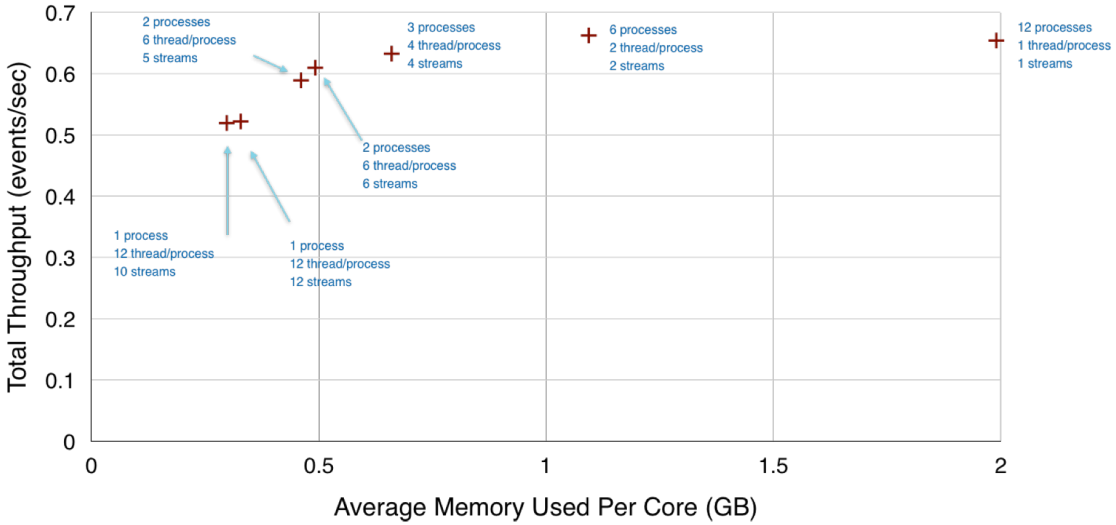
http://science.energy.gov/ascr/research/scidac/exascale-challenges/

- Power. Power, power, power.
  - Naively scaling current supercomputers to exascale would require a dedicated nuclear power plant to operate.
    - "The target is 20-40 MW in 2020 for 1 exaflop."
- High-concurrency/limited memory per thread
- Memory bandwidth
  - "Memory bandwidth is not expected to scale with floating-point performance."
- I/O
  - "The I/O system at all levels – chip to memory, memory to I/O node, I/O node to disk—will be much harder to manage, as I/O bandwidth is unlikely to keep pace with machine speed."
- Many people are skeptical that HEP physicists will be able to address these challenges.

Addressing memory limitations

- Software frameworks form the backbone of HEP software. Fermilab develops two.
  - CMSSW, used by CMS
  - *art*, used by most Intensity Frontier experiments
- While per-process scaling of event processing has worked well until recently, memory constraints are pushing us toward multi-threaded frameworks.

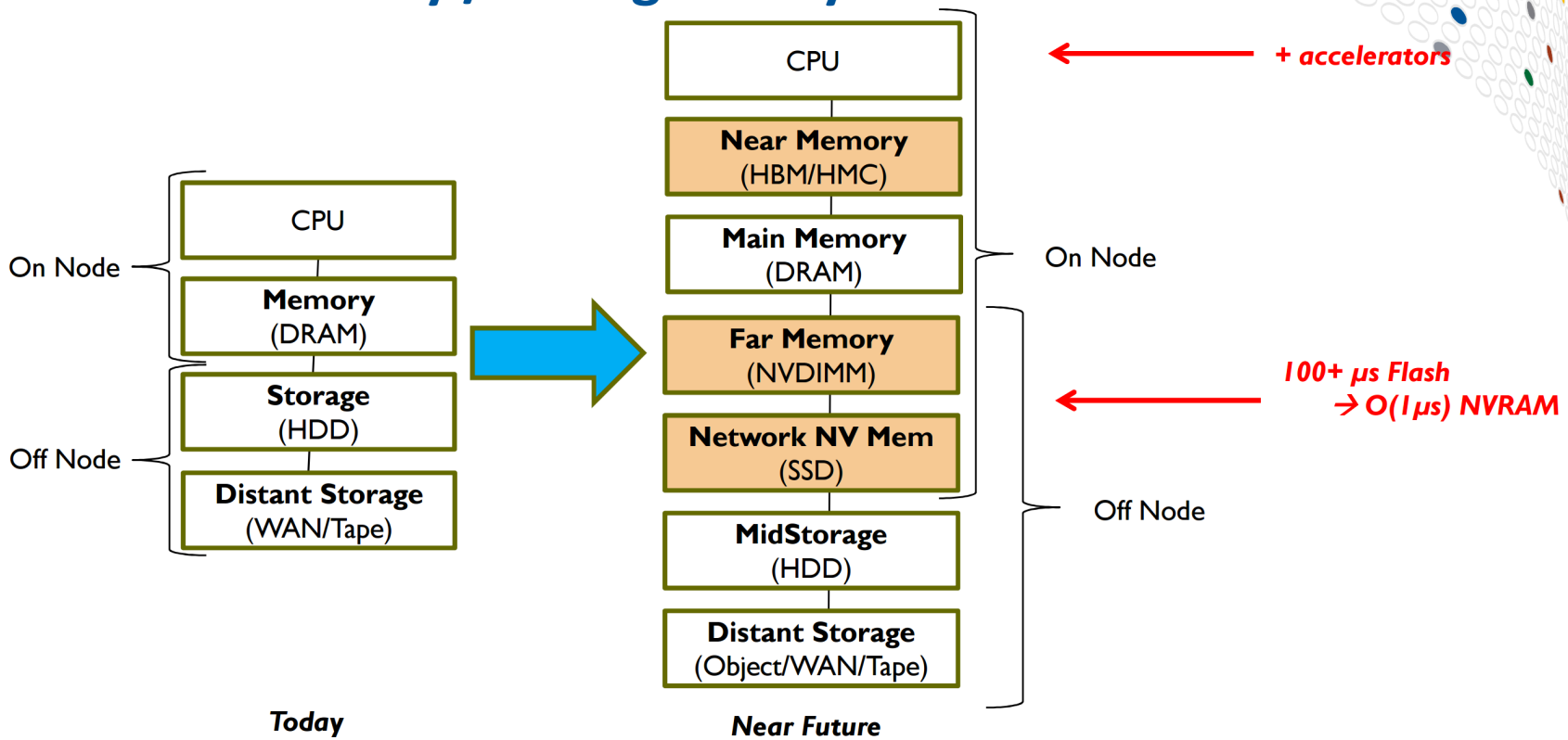
Demonstration of reduced memory through threading in CMSSW. *art* threading work is in progress.



I/O

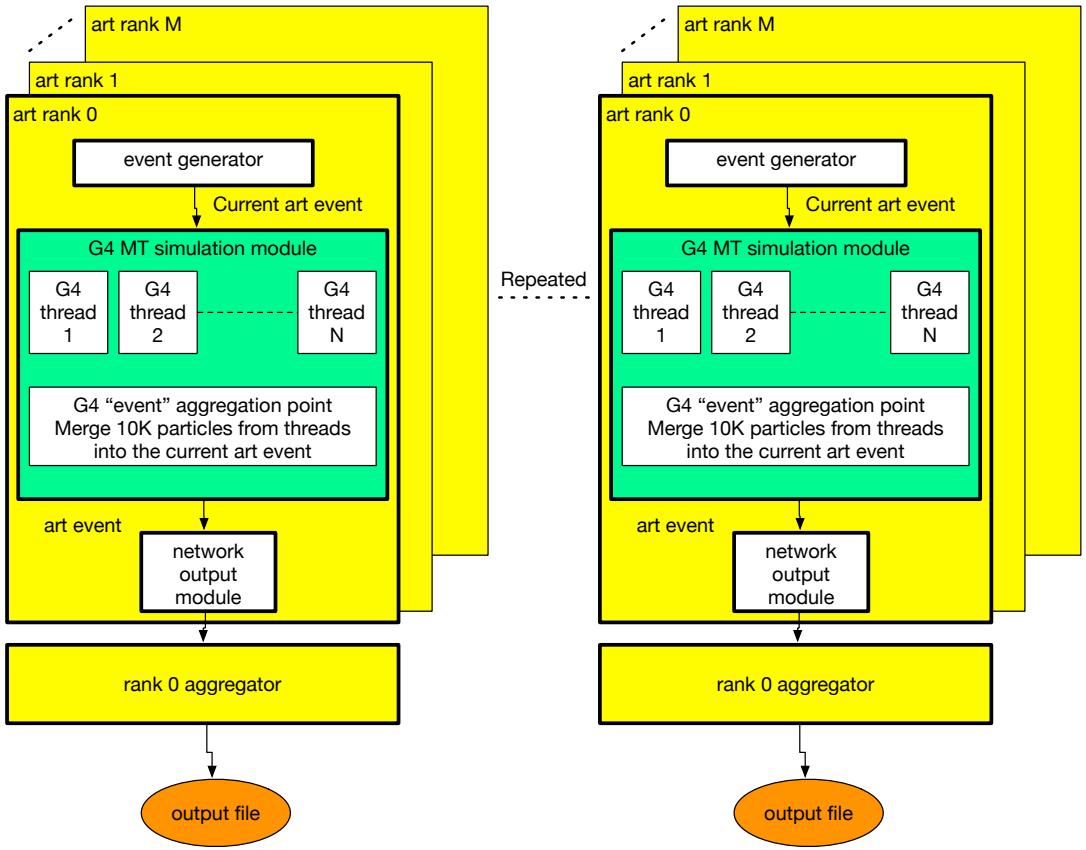
https://hpcuserforum.com/presentations/paris-munich/CRAY.HPCUserForumOctoberParis%202015CRAY.pdf

Trends in the Memory / Storage Subsystem



art-HPC: Addressing I/O trends, et al.

- **Extending the ART Framework to Support Large Scale Multiprocessing for the Intensity Frontier**
  - Partnership with Tom LeCompte at ANL
  - Migration of art to HPC and Mira
  - Using MPI
  - Multi-threaded Geant4
- Target is to produce 10^12 muons for muon g-2 on ALCF Mira
- Architected to addressed
  - limit I/O to filesystem
  - scaling



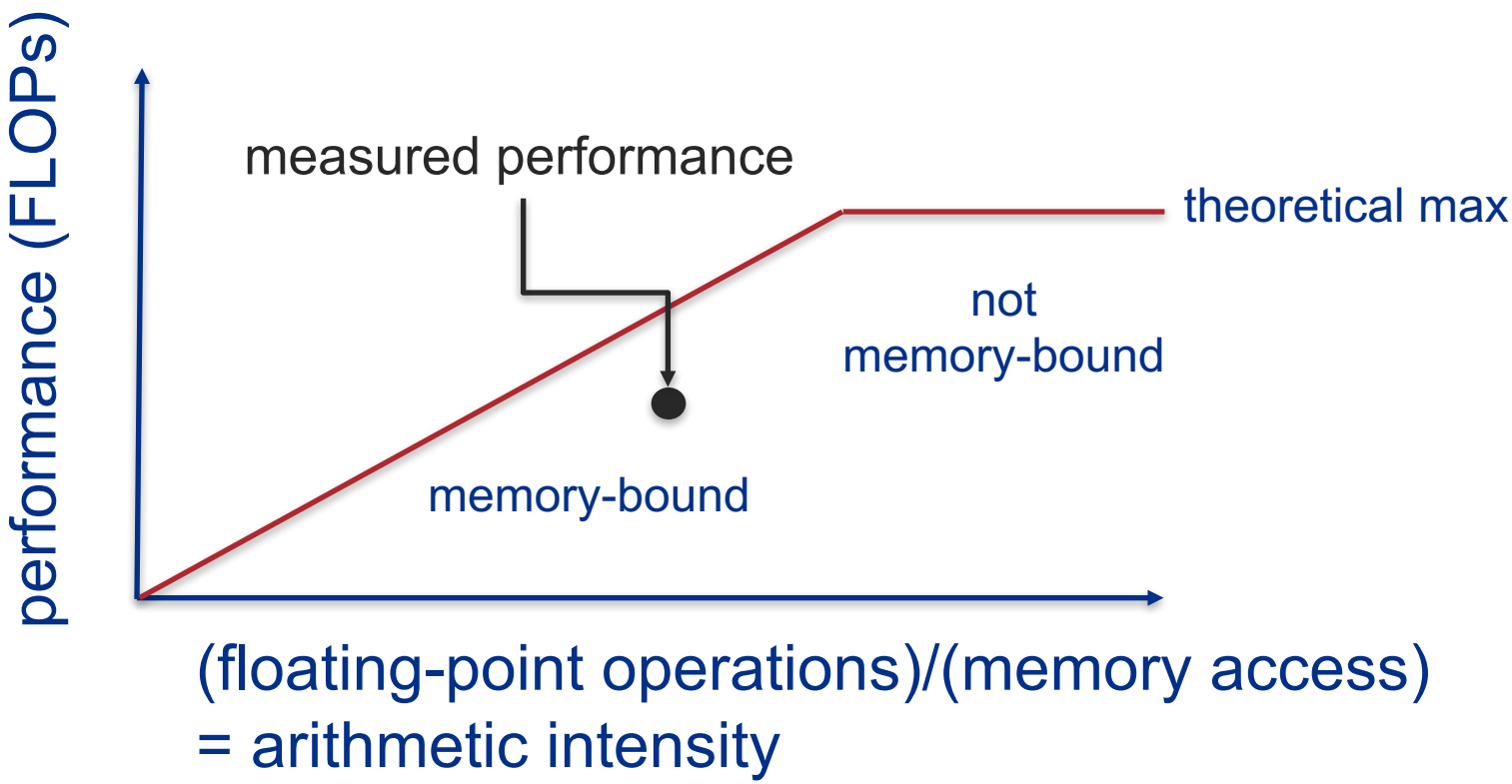
NOTE: Same architecture applied to running a multi-parameter tuning of event generators using collider data analysis on Mira using Pythia

thanks to Jim Kowalkowski

https://cdcvs.fnal.gov/redmine/projects/art-hpc/wiki/

Understanding CPU constraints: the Roofline Model

Basic idea: understand maximum theoretical performance for a given code in order to guide optimization

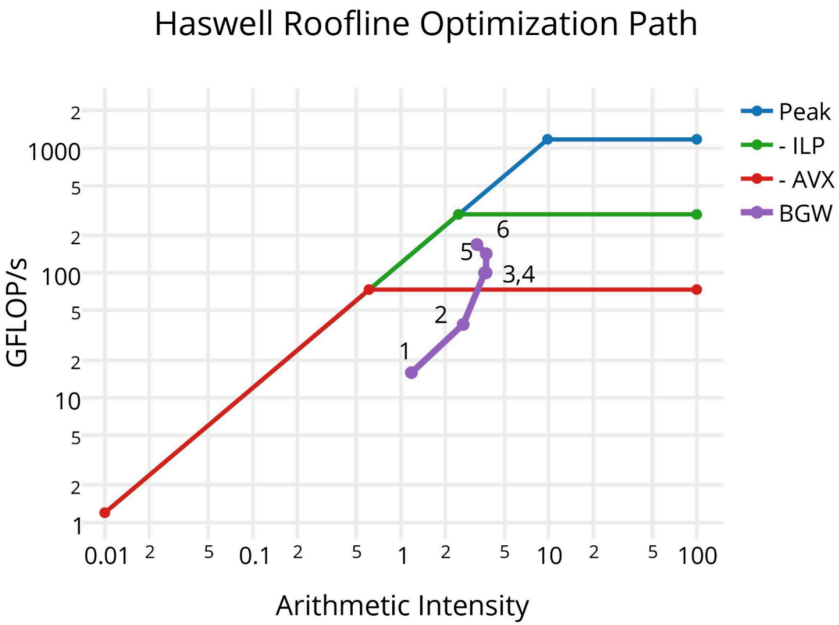


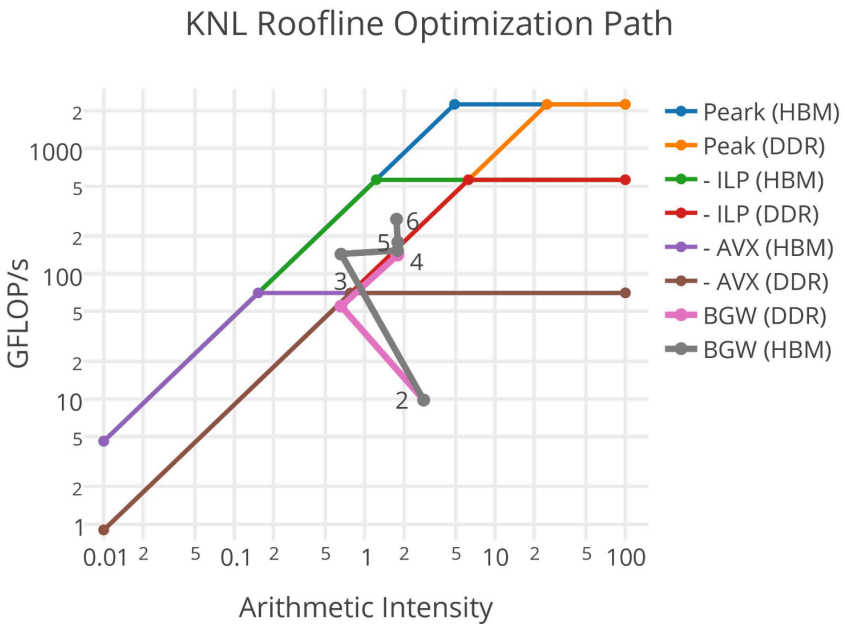
Roofline in the real world

https://anl.app.box.com/v/IXPUG2016-presentation-29

Optimizing the BerkeleyGW code.

The BerkeleyGW Package is a set of computer codes that calculates the quasiparticle properties and the optical responses of a large variety of materials from bulk periodic crystals to nanostructures such as slabs, wires and molecules.





Conclusions

- HEP computing has gone through many epochs
  - BC
  - VAXus Vulgaris
  - Unix Principium
  - Linux Maximus
- We are looking forward to the exascale epoch,
  - Deus Ex Machina
- Exascale computing will require many changes
  - Multithreaded processing
    - In progress
  - I/O changes
    - Just getting started
  - Code optimization
    - Much work to do
    - Can use the roofline model to guide us
- Each computing transition has faced skepticism. We will once again rise to meet the challenges that the new epoch is bringing.