



Science and
Technology
Facilities Council

Future of High Throughput Computing (HTC)

Andrew Sansum

Scientific Computing Department STFC

Head Systems Division

IRIS Technical Director – Joining up STFC's National
eInfrastructure

andrew.sansum@stfc.ac.uk

17/1/20

What is High Throughput (HTC)

- loosely-coupled tasks
- minimal parallel communication requirements
- Focus is on maximum throughput not maximum speed
- Turnaround may be measured in weeks or months
- A workflow may require many thousands of jobs
- Often data processing of long lived data (not work files)
- Data set volumes may be measured in petabytes
- Data may even be stored on tape

Typical HTC use Cases

- Different particle physics events
- different random numbers in a simulations based on Monte Carlo methods
- different model parameters in ensemble simulations or explorations of parameter spaces
- different patient data in large scale biomedical trials
- different parts of a genome or protein sequence in bioinformatics applications



Science and
Technology
Facilities Council

Historical Context

Lets have a look at the past



First Encounters with HTC

Undergraduate Days



1978 – Punched cards

Postgrad



Interactive job prep



1981 - GEC 4085

```
Update COBUGG : Trunc Xlate Top Nonum Nulls Asis -----
Cmd => ----- Scope 01,72 Scr
..... -1-+-----2-----3-----4-----5-----6-----
000100 //HELLO JOB (001), 'COMPILE/RUN HELLO', CLASS=A,MSGCLASS=X
000200 //COB EXEC PGM=IKFCBL00,REGION=4096K,
000300 // PARM='LIST,LOAD,SIZE=2048K,BUF=1024K'
000400 //STEPLIB DD DSN=SYSC.LINKLIB,DISP=SHR
000500 //SYSPRINT DD SYSOUT=*
000600 //SYSUT1 DD UNIT=SYSDA,SPACE=(460,(700,100))
000700 //SYSUT2 DD UNIT=SYSDA,SPACE=(460,(700,100))
000800 //SYSUT3 DD UNIT=SYSDA,SPACE=(460,(700,100))
000900 //SYSUT4 DD UNIT=SYSDA,SPACE=(460,(700,100))
001000 //SYSLIN DD DSN=&LOADSET,DISP=(MOD,PASS),
001100 // UNIT=SYSDA,SPACE=(80,(500,100))
001200 //SYSIN DD DSN=HMVS01.SOURCE(HELLO),DISP=SHR
001300 //GO EXEC PGM=LOADER,PARM='MAP,LET',COND=(5,LT,COB),REGION=100K
001400 //SYSLIN DD DSN=*.COB.SYSLIN,DISP=(OLD,DELETE)
001500 //SYSLOUT DD SYSOUT=*
001600 //SYSLIB DD DSN=SYSC.COBLIB,DISP=SHR
001700 //SYSOUT DD SYSOUT=*
001800 //
***** End of data *****
```

MVT JCL

Remote Submission
HASP/Mast

Lineprinter
Output



IBM 360/195

Definitional Criteria for a Distributed Processing System

Philip Enslow, *"What is a Distributed Data Processing System?"* Computer, January 1978

Proposed Definition

- Multiplicity of resources
- Component interconnection
- Unity of control
- System transparency
- Component autonomy

Perceived Benefits

- High Availability and Reliability
- High System Performance
- Ease of Modular and Incremental Growth
- Automatic Load and Resource Sharing
- Good Response to Temporary Overloads
- Easy Expansion in Capacity and/or Function

Intel – 1991

The i860™ XP Second Generation of the i860™ Supercomputing Microprocessor Family

Target Markets:

- *Massively Parallel Supercomputer and Multi-Processing Systems*
- *Super Workstation & servers*
- *High End Workstation Graphics/Accelerator Sub-systems*

High Throughput Computing Performance

- *"Number Crunching" Floating-Point Capability*
- *RealTime 3D Graphics Visualization*

Particle Physics experimenting with x86 in 1990 to provide “High Computational Throughput”

Resource Scheduling - Condor

Mechanisms for High Throughput Computing (1997)

*“Floating point operations per second (FLOPS) has been the yardstick used by most High Performance Computing (HPC) efforts to rank their systems. Little attention has been devoted by the computing community to environments that can deliver large amounts of processing capacity over very long periods of time. We refer to such environments as **High Throughput Computing (HTC) environments**”*

M. Livny, J. Basney, R. Raman, and T. Tannenbaum, Department of Computer Sciences, University of Wisconsin. May 9 1997

CONDOR Team “working for more than a decade to provide High Throughput Computing tools”

Beowulf Clusters (1994-1998)

The *first Beowulf* cluster was *built* by Donald Becker and *Thomas Sterling* at NASA's Center for Excellence in Space Data and Information Sciences in *1994*. ... was to *build* Commodity Off-The-Shelf (COTS) based systems.

Beowulf: A Parallel Workstation For Scientific Computation (1995)

Proceedings of the 24th International Conference on Parallel Processing. Stirling et al.

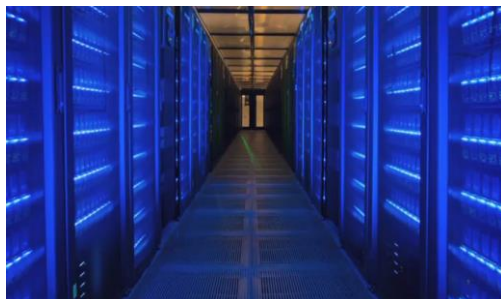
The Beowulf evolved for parallel applications but in turn the HTC community set about converting their RISC based clusters to **Commodity** off the Shelf Linux Clusters

High Throughput Computing - 1999

Early effort to exploit Linux at RAL for HTC workloads. May 1999

The **Central Simulation Facility (CSF)**. Dual Pentium 450 Cluster for particle physics

Present day

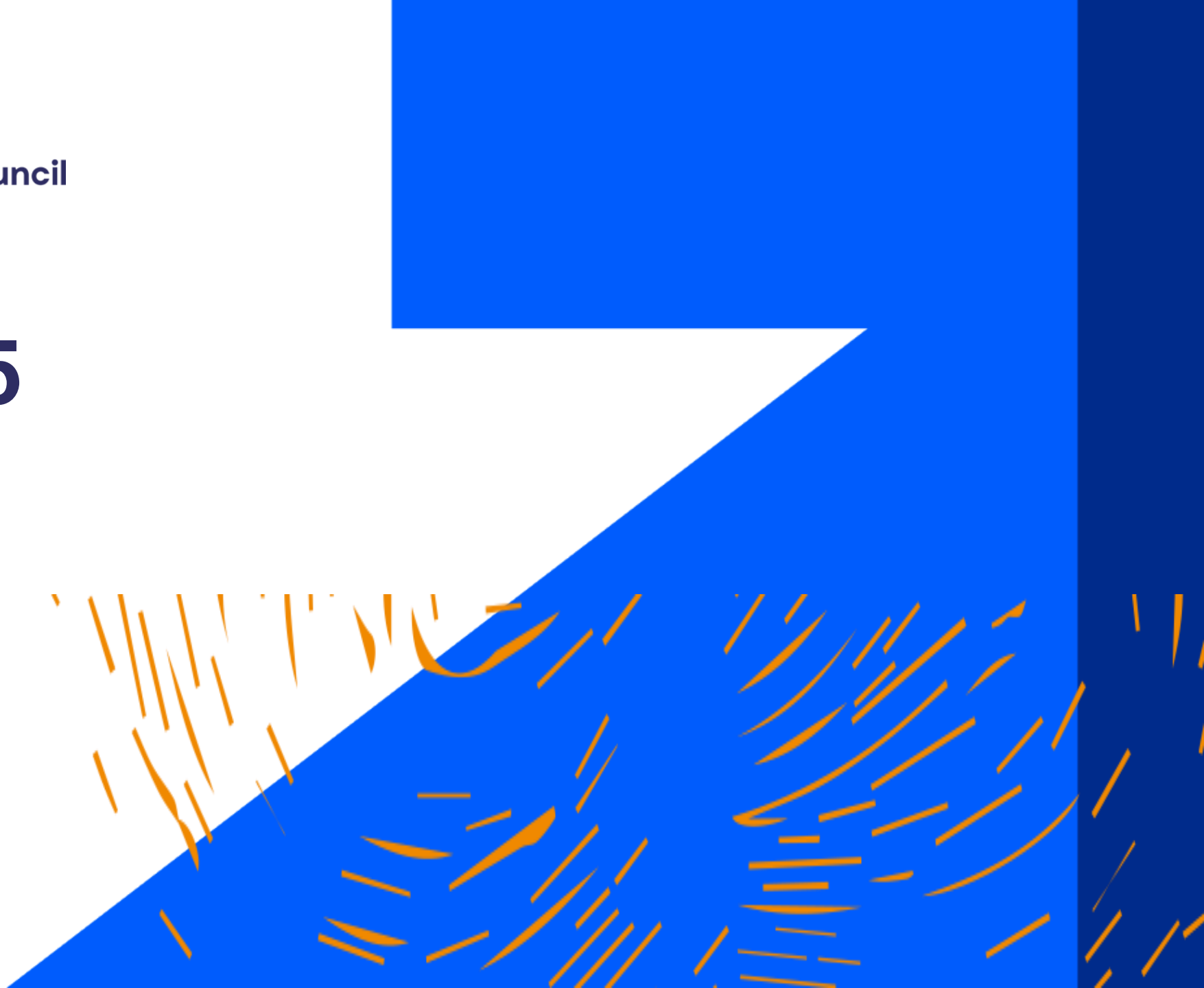




Science and
Technology
Facilities Council

HTC 2025

And Forward



Predicting the Future Not Easy

2025 Not so far off. CPU Procurements made in FY20 will deliver in Q1 2021 and if assume 6 year life will phase out in Q1 2027. "Analyst firms have about a 10% accuracy rate predicting market trends 24 months into the future (0.4 probability)"
Found on the internet – absolutely no justification

Trying to predict the future is like trying to drive down a country road at night with no lights while looking out the back window.

Peter Drucker

f t v +

Future Looking Night Drive



2007

Easier – to talk about the past than the future. What pieces have we already got in play – probably will still be there in 2026!

Is That A Pita In Your Pocket?

I've got a cell phone, a pocket organizer, a beeper, a calculator, a digital camera, a pocket tape recorder, a music player, and somewhere around here, I used to have a color television. Sometime in the next few years, all of those devices are going to meld into one. It will be a box less than an inch thick and smaller than a deck of cards. (The size will be determined by what's convenient to hold, not by the technology inside.) The box will have a high-res color screen, a microphone, a plug for a headset or earphones, a camera lens, wireless connectivity, cell phone and beeper functions, a television and radio receiver, a digital recorder, and it will have enough processing power and memory to function as a desktop system. It will be able to dock with a keyboard and full-size monitor. Oh yes, and it will handle e-mail, as well.

Most important of all, it will have both speech recognition and speech synthesis. It will listen and respond in English or whatever language you need, and yes, it will be a translator, too. It will be an agent, going out and doing cyber-errands for you. For instance: I need a Japanese restaurant in Tulsa, near the Ramada Inn. Book a reservation and arrange transportation. If there's no Japanese restaurant, try for Italian. Or, voice-mail Bob as follows: "Bob, we accept your offer, but we'll need a draft of the deal memo by the 15th. Let me know if that's a problem."

I call this device a Personal Information Telecommunications Agent, or Pita for short. The acronym also can stand for Pain In The Ass, which it is equally likely to be, because having all that connectivity is going to destroy what's left of everyone's privacy.

David Gerrold is a Hugo and Nebula award-winning author who writes about computing. Visit his Web site at www.gerrold.com.

Some interface gurus: Science-fiction writer David Gerrold (top); Jim Spohrer, Ph.D., senior manager of computer science, IBM Almaden Research Center (left).

Smart Reseller: www.smartreseller.com December 20, 1999 61

1999

Existing Themes – What do we have

- Moore's law is slowing – but demand continues to rise
- True commodity hardware is far behind us – we fell into the gap
- Sweating the hardware capabilities – no magic solution
- Handling/exploiting many core
- Some communities are only now discovering batch – this is OK!
- The convergence of virtual research environments (VRE) and prompt response

Existing Themes – (II)

- Once you grow beyond capacity of a single site - more communities need access to large HTC eInfrastructures
- Federation service such as AAI become vital
- Workflow and data management systems are increasingly necessary
- Many workflows are International
- Data placement models will evolve
- The commercial cloud remains an economic challenge to

exploit



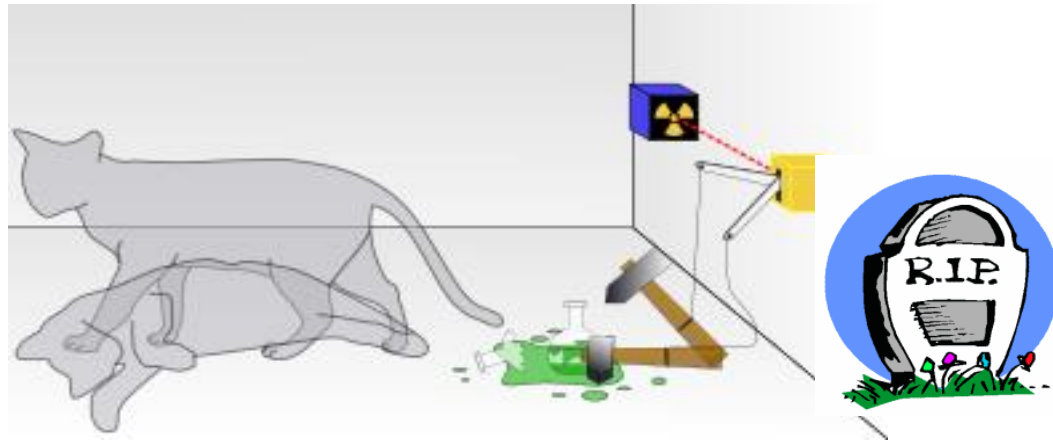
Science and
Technology
Facilities Council

Lets Start with Hardware



Health of Moore's law

“Debate Over Health Of Moore’s Law Continues ... At Semicon West 2019, CEOs from across the industry continue to debate whether Moore’s Law is alive or dead.”



“it’s **completely** alive is because right now we’re facing another decade or two of amazing opportunities that themselves economically will drive the push for technology without a stop.

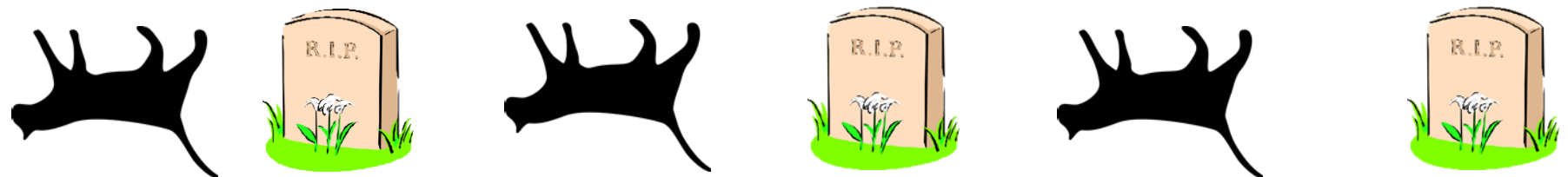
Maybe it’s **not exactly the same curve** that Moore actually drew, it doesn’t matter. The impact is what matters of the exponential.”

Aart de Geus Semicon 2019

“The way to think about it is Moore’s Law is the behaviour of an exponential that has techonomic feedback on the exponential that drove a revolution of what mankind can do. “

it doesn’t [anymore] deliver simultaneous improvements in power, performance, area and cost.

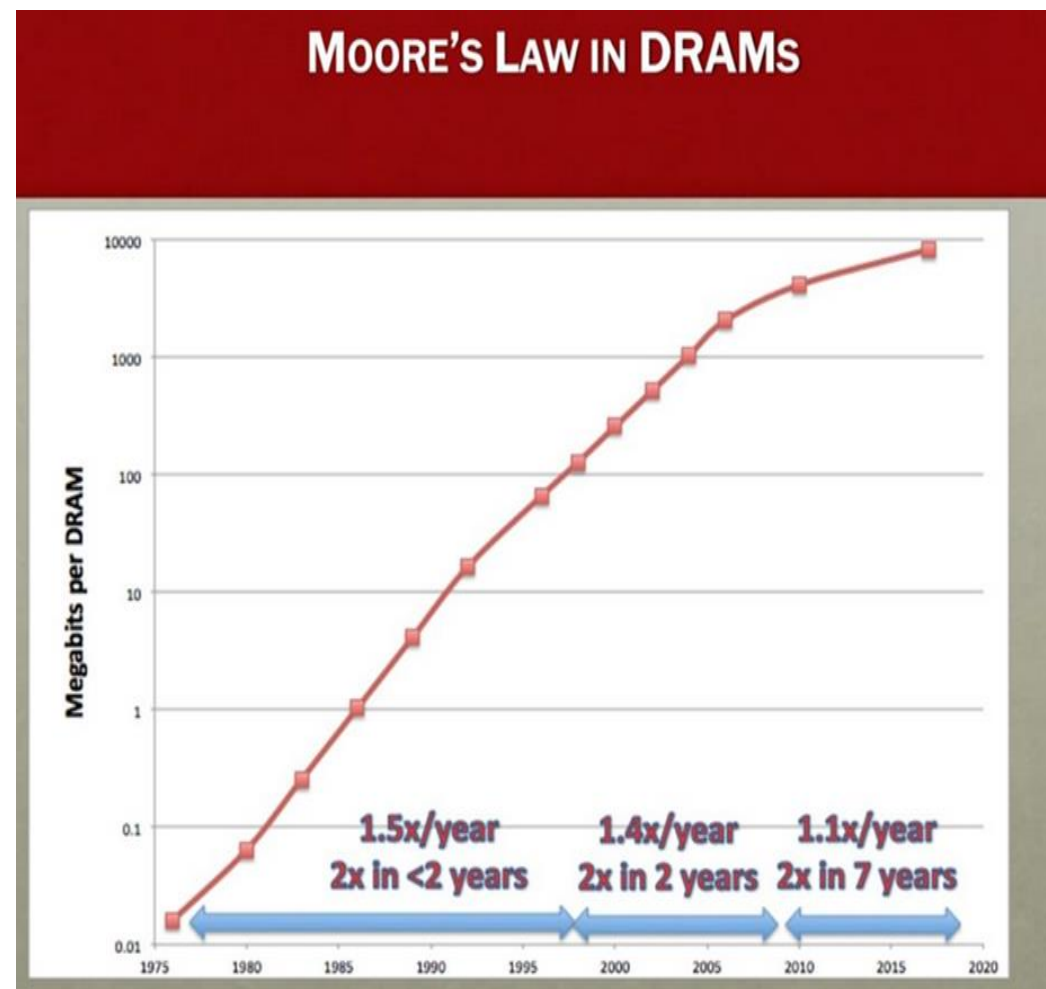
Gary Dickerson, CEO of Applied Materials



Memory (Semicon West 2019)

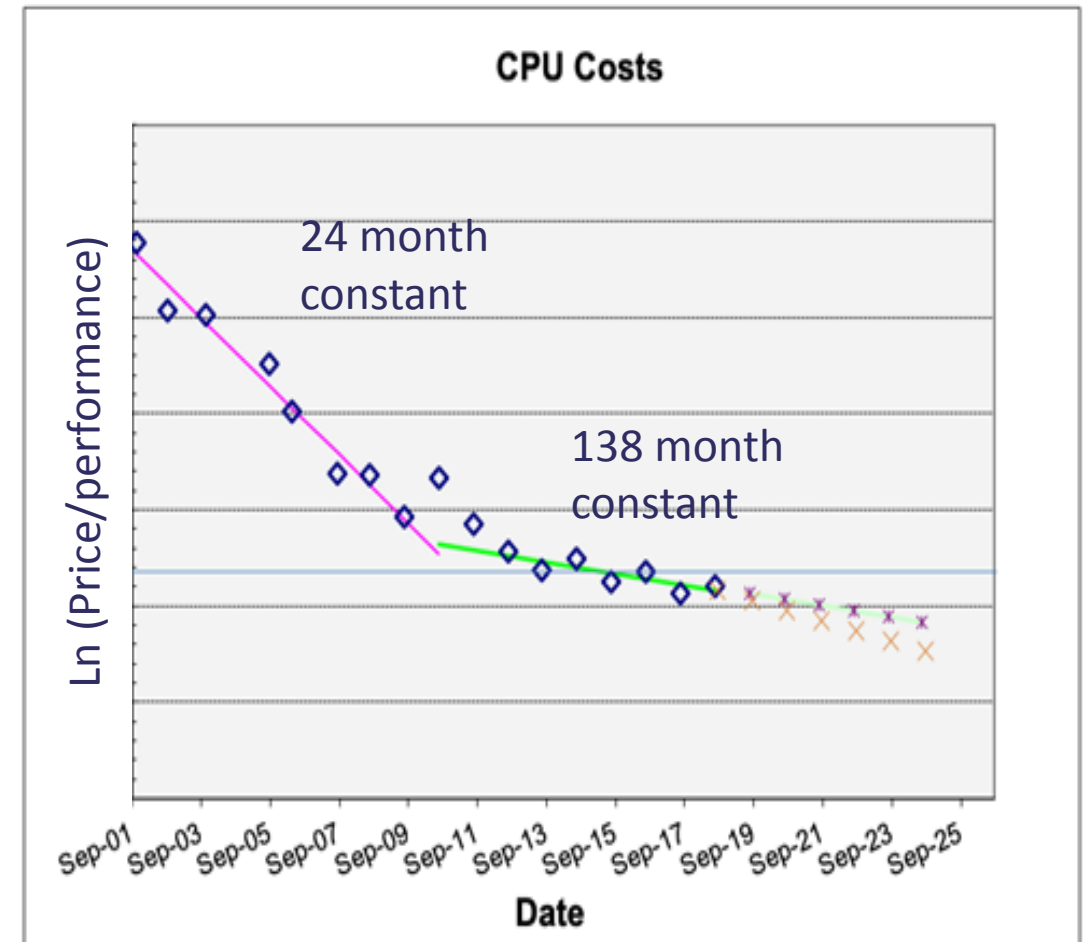
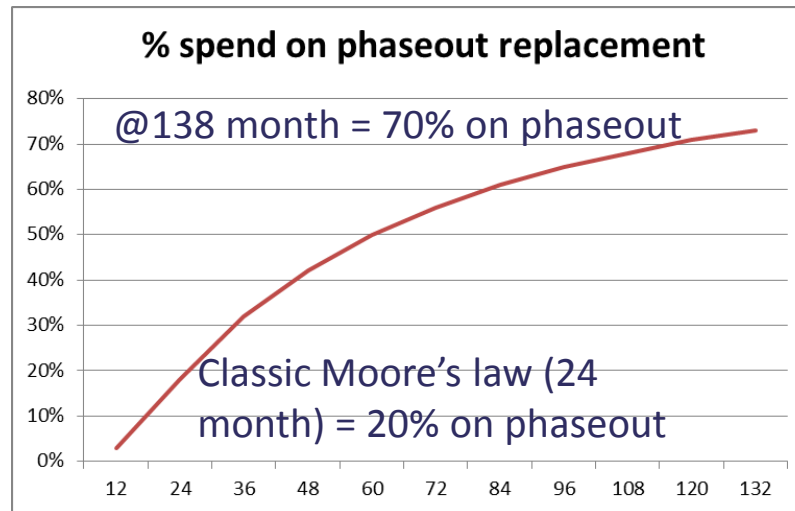
For almost a decade, Moore's Law has slowed down significantly "A lot of innovations have been driven in memory and technology, for example, going from 2D NAND to 3D NAND But there is no question that Moore's Law is significantly challenged in memory and storage. Looking at 10 years ago versus today in NAND as well as in DRAM, the year/year bit growth that you could get from one technology transition to the next technology transition, that bit growth, that cost decline capability has more than halved now, and certainly there are challenges."

Sanjay Mehrotra, CEO of Micron Technology



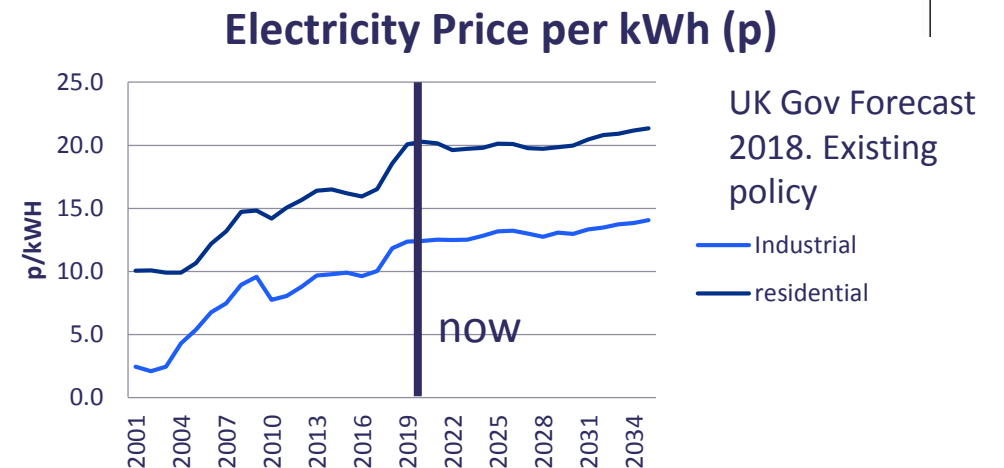
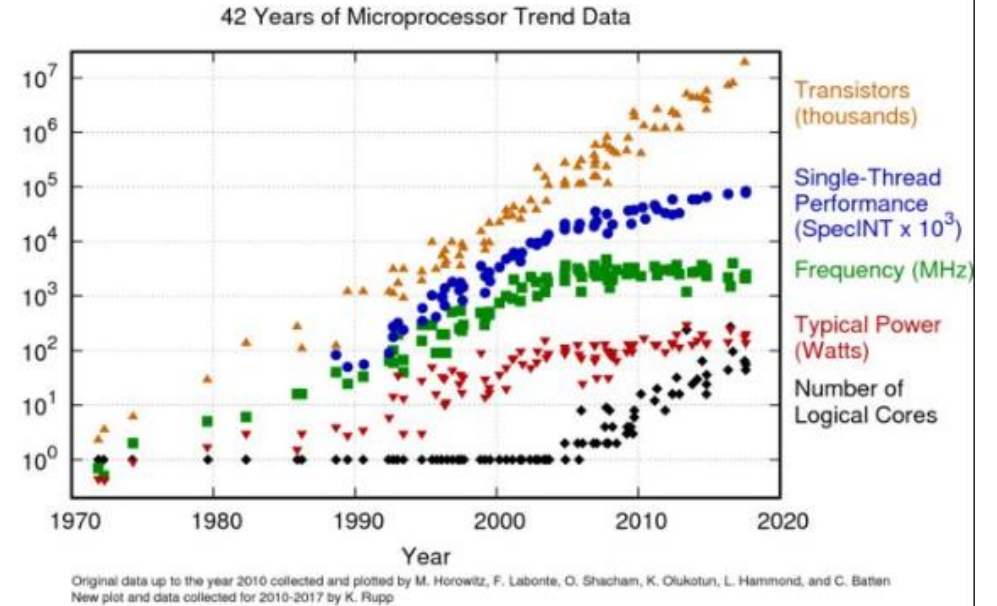
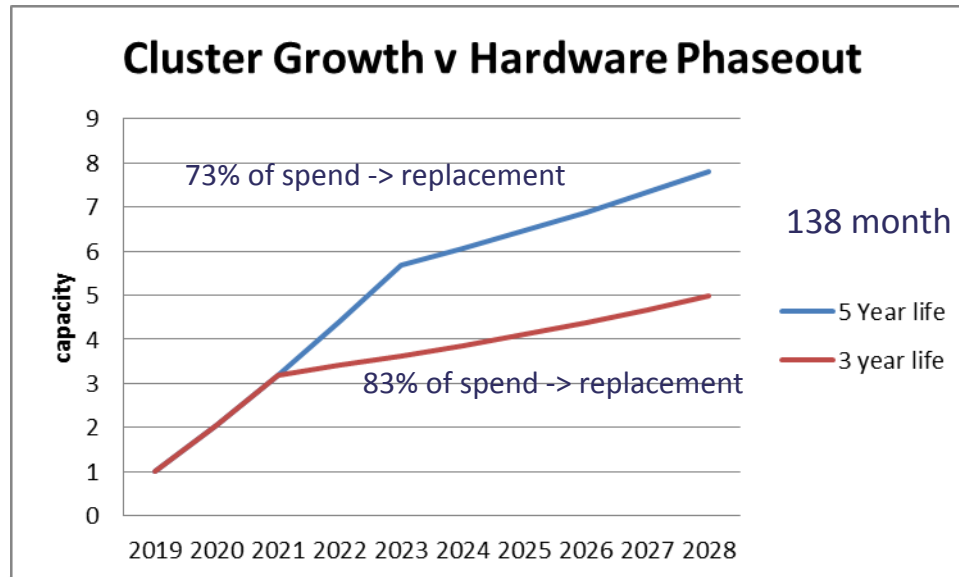
Price per unit Performance

- For HTC its “bang per buck” that matters
- HTC sites rarely buy the newest and fastest – looking for best value (biggest volume)
- Flat cash – weak Moore’s Law– **steady state majority of investment goes into phase-out replacement**

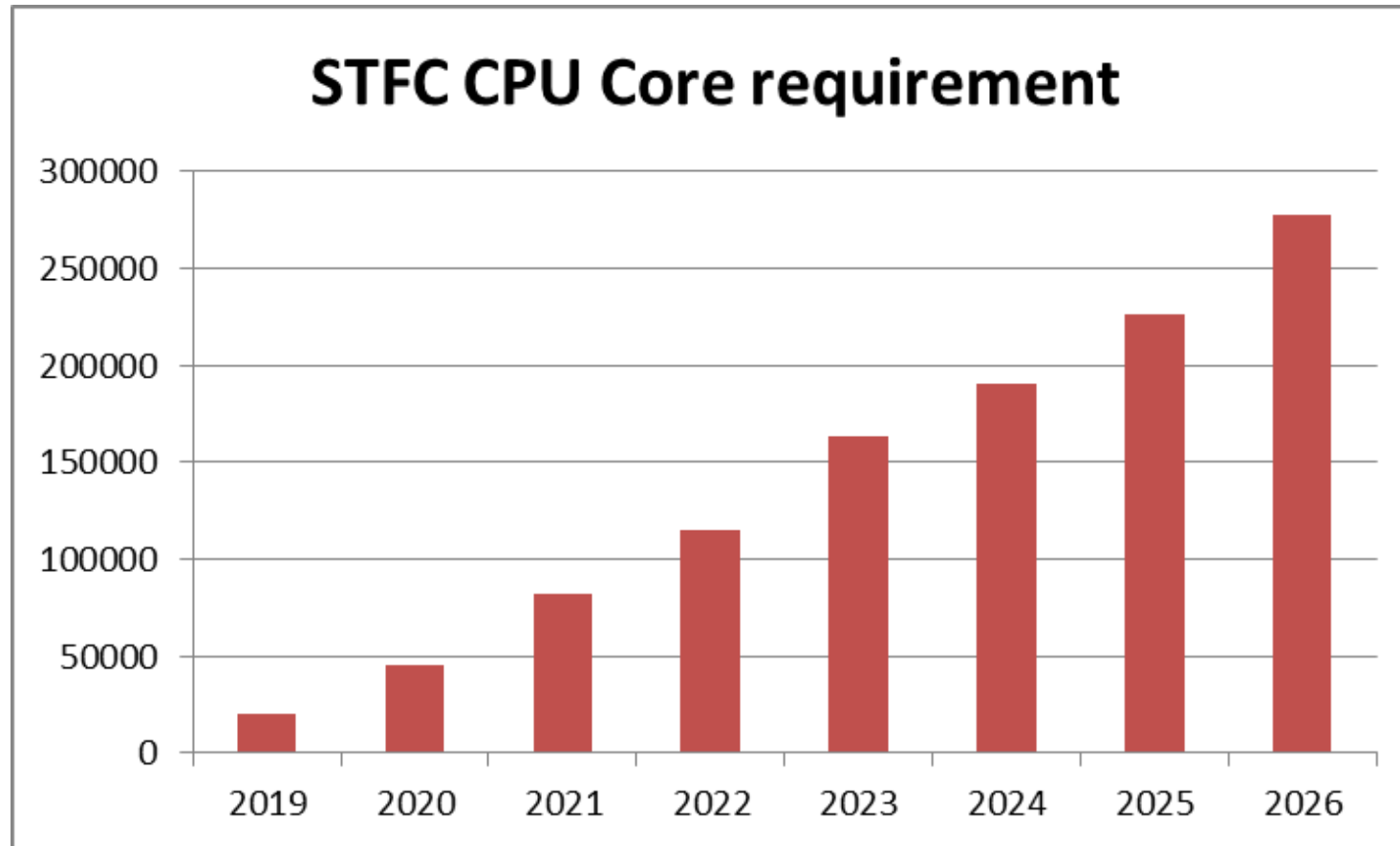


Electricity Consumption

- Power consumption of CPU limiting performance
- Prices forecast to be relatively flat after recent rises
- Will be increased pressure to reduce CO2 footprint
- Can apply energy efficiency constraint for procurement – may raise hardware costs
- Or by earlier replacement – but at the cost of cluster growth rate



STFC Compute Requirements



ARM – not going to help ... much

- Limited benchmarking in science community – but see:

“Evaluating the Arm Ecosystem for High Performance Computing – Jackson et al. EPCC - 2019” based on ThunderX2 – ARMV8

“we have also demonstrated that applications can achieve similar, or better, performance on such a system when compared with a range of existing HPC system architectures.....a viable alternative”

- For x86 servers CPU represents 30-50% of total server cost, but ARM server class CPUs at comparable price. Isn't going to save us.
- Benefits may rather accrue from increasing competition in the market rather than miracle CPU.
- **May Help Power Consumption**

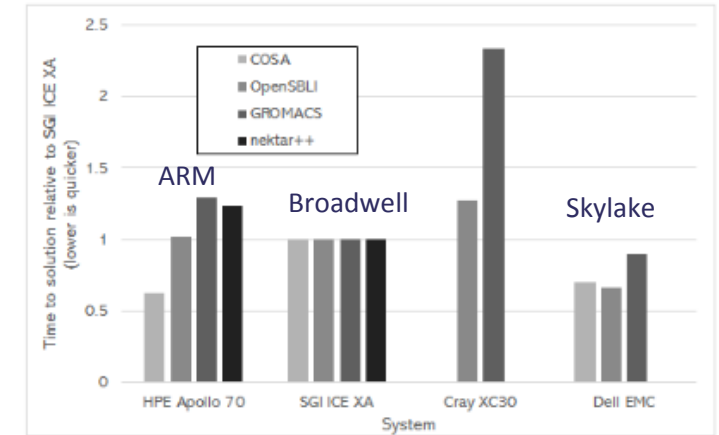
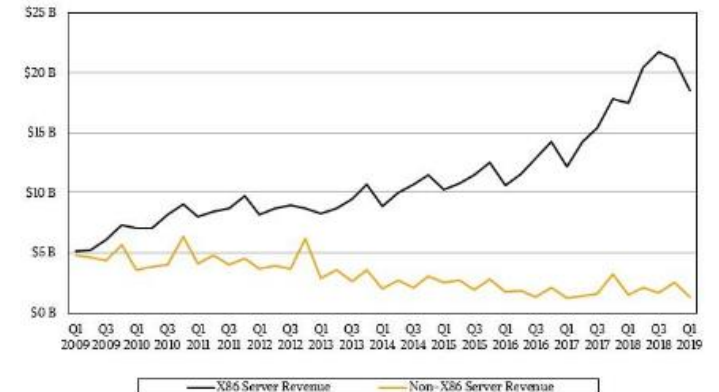


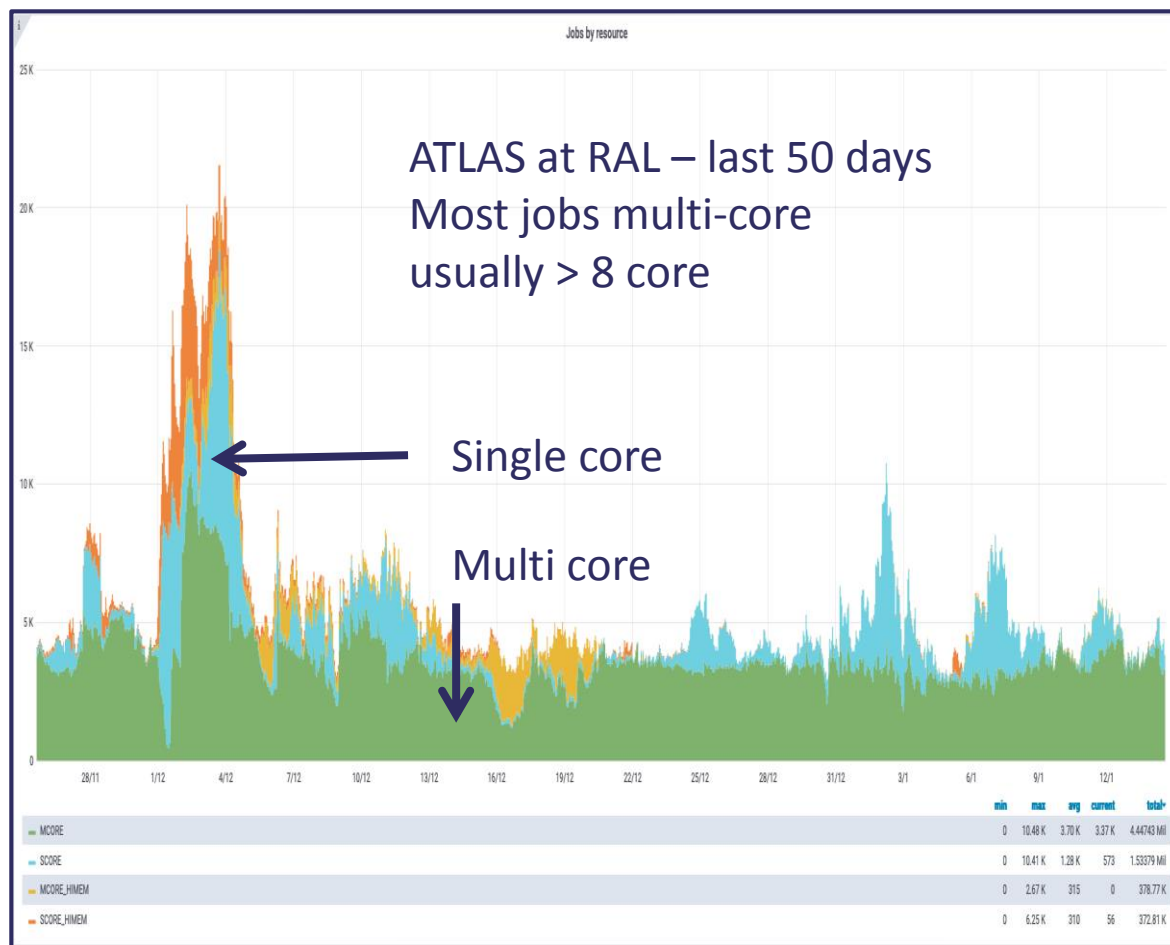
Figure 1: Comparison of application time to solution using 16 nodes normalised to the SGI ICE XA system. Values lower than one are faster than the SGI ICE XA system.



X86 server v non x86 server market share

HTC but Exploiting Many Core

- As system core count increases so too does complexity of cluster workload manager packing jobs with different memory requirements.
- User pilot jobs can instead schedule separate threads to optimise memory footprint.
- If done badly can lead to cluster inefficiency
- Done well simplifies cluster scheduling



Sweating Hardware: Code and Algorithms

Most Gains to be Made Here – Do it Smarter

- Architecture specific compilation
- Adapt to many hardware types - heterogeneous hardware – validation nightmare
- Code re-engineering to exploit vector and other hardware units
- Algorithmic improvement
- Paradigm shift – eg ML rather than brute force



Science and
Technology
Facilities Council

eInfrastructure



Modern (Mature) Descriptions

“High-throughput computing (HTC) is a powerful paradigm that allows vast amounts of independent work to be performed simultaneously across many loosely coupled computers. HTC aims at integrating multiple computing systems to enable large numbers of computing tasks to be schedule and completed as quickly as possible.”

International Journal of Trend in Research and Development, Volume 5(4), ISSN: 2394-9333 2018

- Single Cluster
- Homogeneous Multi-Cluster or federated cloud
- Federated eInfrastructure – eg Grid
- Opportunistic use – eg Commercial cloud or spare cycles – eg on HTC

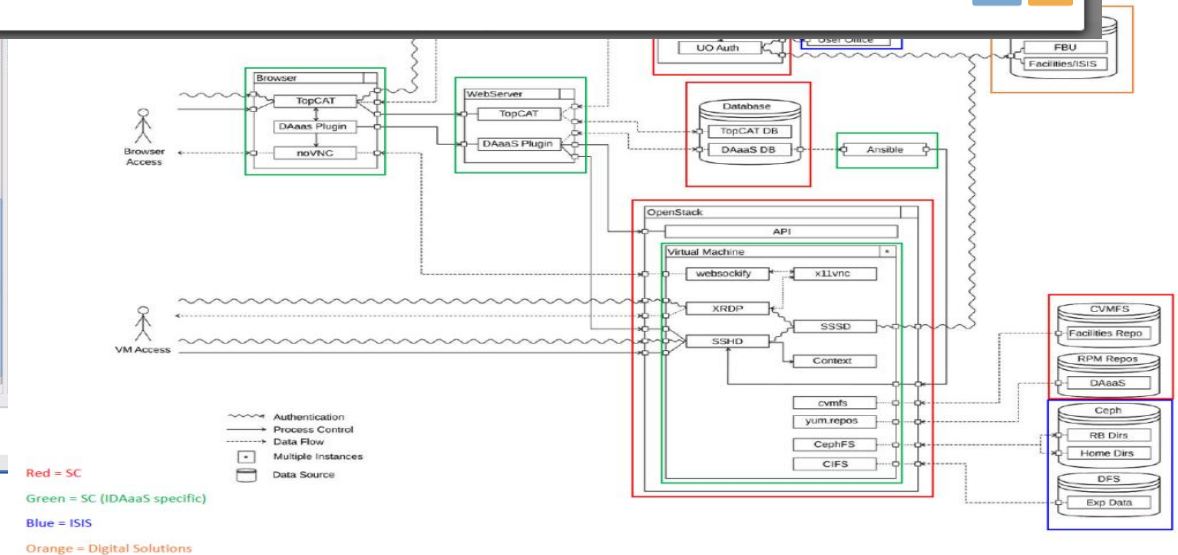
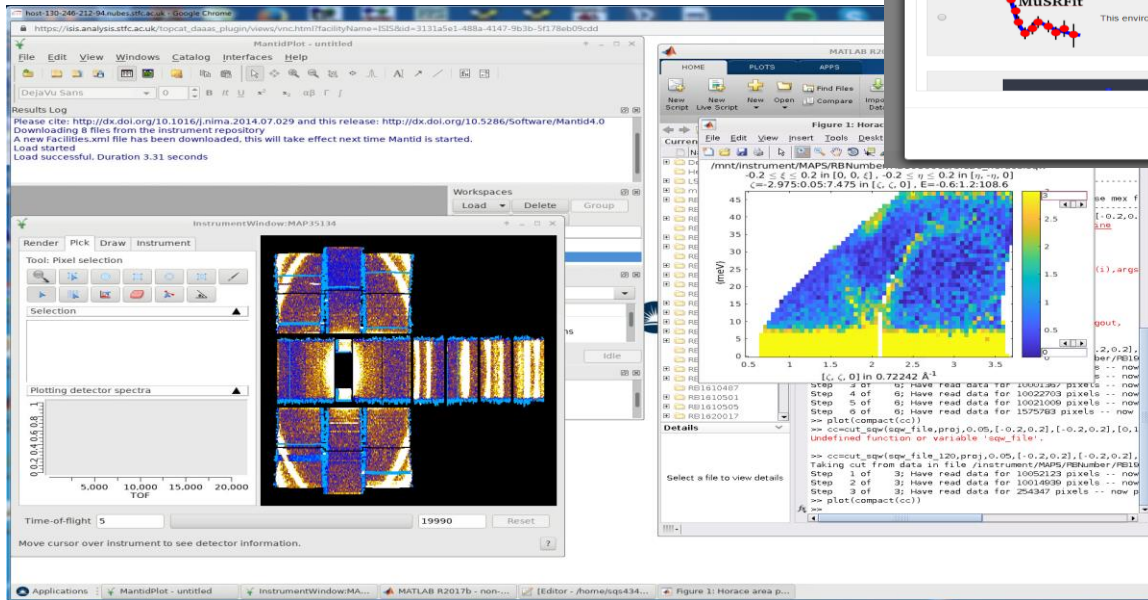
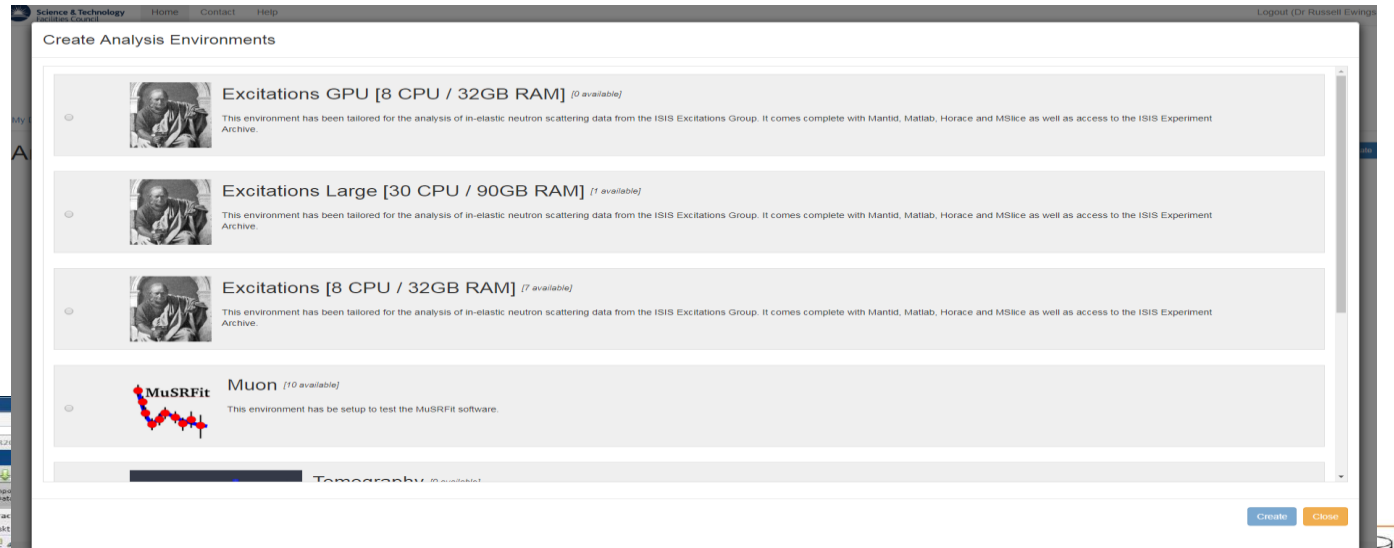
Traditional Batch Compute Lives!

- While some community requirements have grown beyond single clusters, some are only now growing into them.
- In the dash to join up our eInfrastructure we continue to need classical batch services.
- Login and submit some jobs.
- Usually department level “interactive” services
- But environment typically “one size fits all”

Virtual Research Environments for Science: The IDAaaS system

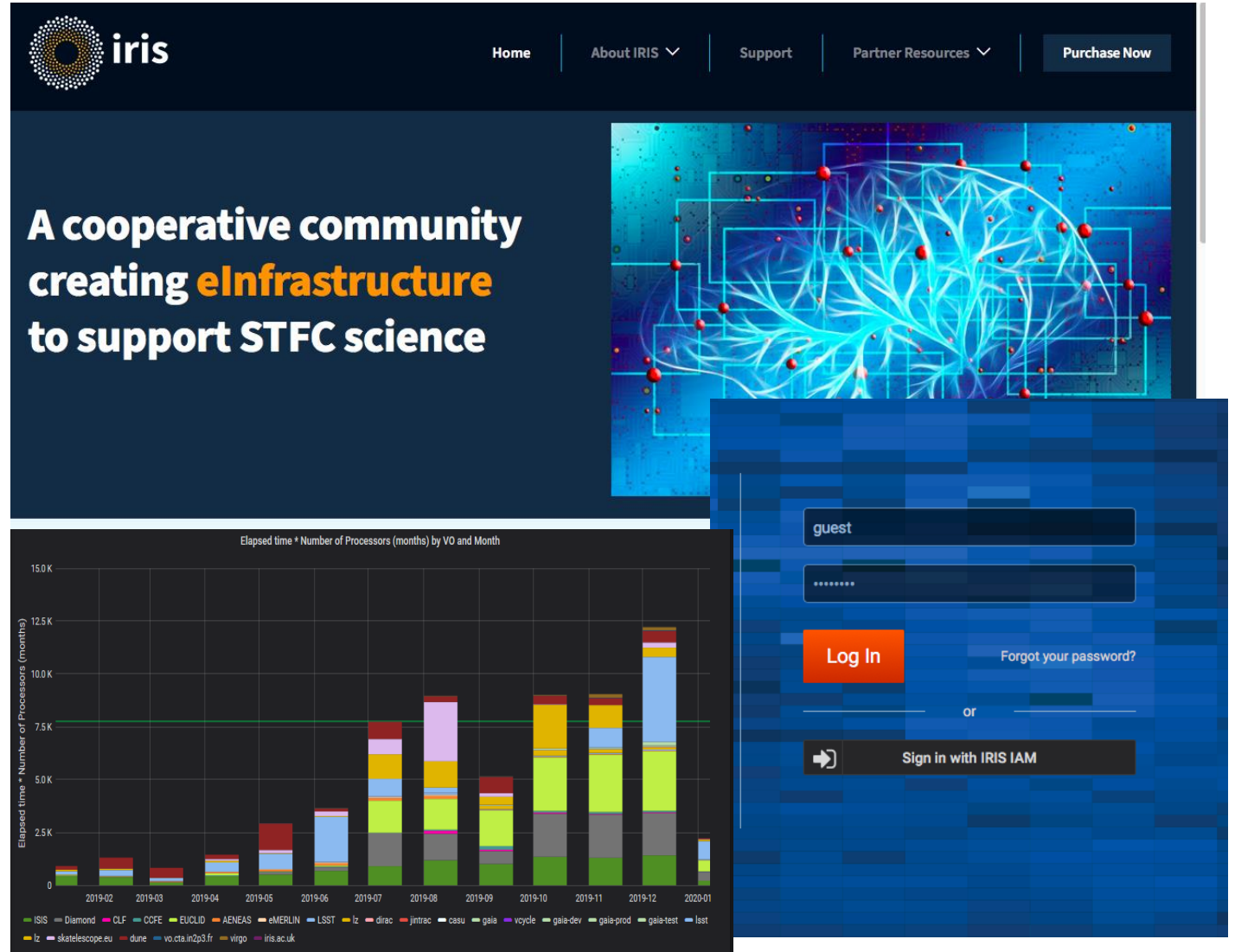
What it is: ISIS Data Analysis as a Service

- VMs Tailored to community
- Dynamic Creation
- Expands with demand
- OpenStack Platform
- Can exploited multi-site
- Burst capacity



Federating STFC Distributed Computing

- STFC has diverse compute infrastructure deployed around UK
 - Single platform
 - Multi-site
 - Many user communities
- IRIS – Capital only project deploying hardware. Coordinating STFC Compute Community, Resource Sharing

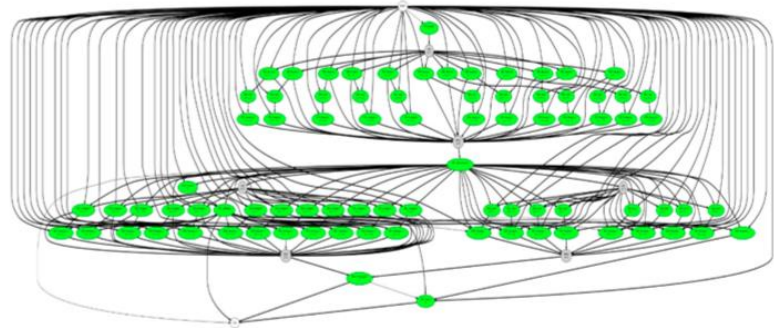


The screenshot displays the IRIS website interface. At the top, the IRIS logo is visible alongside navigation links for Home, About IRIS, Support, Partner Resources, and Purchase Now. The main banner features the text: "A cooperative community creating **eInfrastructure** to support STFC science". Below the banner, a bar chart titled "Elapsed time * Number of Processors (months) by VO and Month" shows data from February 2019 to January 2020. The y-axis represents the product of elapsed time and the number of processors in months, ranging from 0 to 150K. The x-axis lists months from 2019-02 to 2020-01. The chart is a stacked bar chart with various colors representing different Virtual Organizations (VOs). A legend at the bottom identifies the VOs: ISIS, Diamond, CLF, CCFE, EUCLID, AENEAS, eMERLIN, LSST, lz, dirac, jinrac, casu, gale, vcycle, gale-dev, gale-prod, gale-test, lsst, skatelescope.eu, dune, vo.cta.in2p3.fr, virgo, and iris.ac.uk. To the right of the chart, a login form is visible with fields for "guest" and a password, a "Log In" button, a "Forgot your password?" link, and a "Sign in with IRIS IAM" button.

Federated – multi-site Slurm on OpenStack

Euclid's Compute Requirements

- First IRIS runs summer 2018
- Data-flow application model
- Uses cluster filesystem
- Simulation run can take 150,000 core hours
- IRIS resource reservation at multiple sites
- Limited options for cluster filesystem



- Euclid Federating Multiple SLURM instances on STFC's IRIS Infrastructure

StackHPC

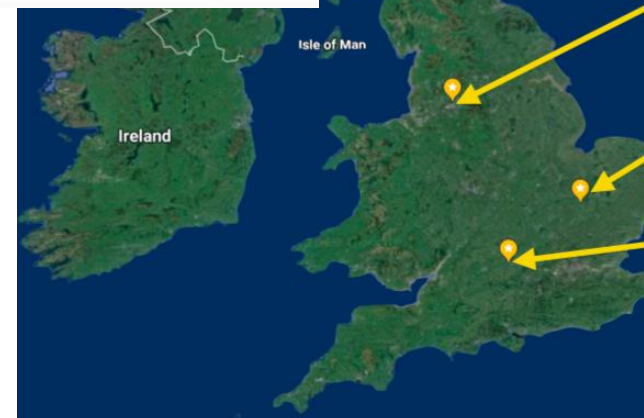
Federated Compute Platform

Royal Observatory, Edinburgh
2 VMs plus long-term storage

University of Manchester
Work in progress

University of Cambridge
39 VMs, 1026 vCPUs, 5.8TB

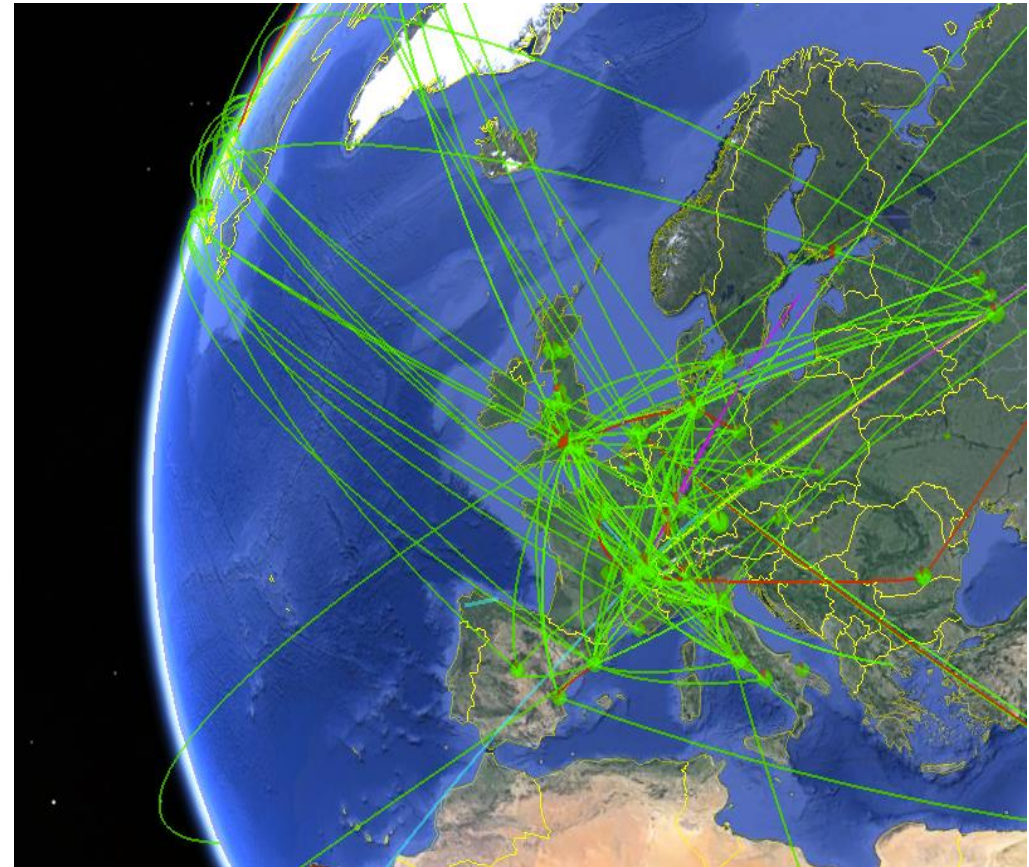
Rutherford Appleton Laboratory
190 VMs, 1728 vCPUs, 29.3TB



What about the Grid?

- *Coordinated resource sharing and problem solving in dynamic, multi-institutional virtual organizations.*
The Anatomy of the Grid - Foster, Kesselman and Tuecke 2001
- **Sounds exactly like what we need!**
- Works well for a few large LHC communities.
- Will still be going strong in 2025 and will probably have a few more large user communities using it.
- Challenging for small communities who do not need largest possible scale

LHC Computing Grid – 900,000 cores



Commercial Cloud

- Already routinely exploited by some STFC communities
- Best fit for:
 - short term (eg burst)
 - rapid deployment of diverse services
 - low data volume
- Not yet compelling price / convenience
- Still requires expertise in deploying service framework
- Challenges of vendor lock in of data
- Main obstacles are not technical but financial / legal detailed in UKRI Roadmap Document for cloud

Predictions → 2025+

- CPU performance gains will continue but (on average slowing)
- More communities will exploit federated eInfrastructure
- VREs will be increasingly exploited
- Data will become increasingly federated
- Workflow management systems will be increasingly necessary
- Heterogeneous hardware capabilities will be increasingly exploited
- Federated eInfrastructure will be increasingly vital
- New user communities will be exploiting the grid paradigm
- Convergence of interactive and pleasingly parallel

Final Observations

- Many science communities are currently constrained by the IT infrastructure available
- For the very very largest – there will be new technical challenges to deliver sufficient compute and storage.
- For some – brute forcing the compute isn't going to deliver the needed performance gain – they need to get smarter
- For most communities however – the solutions are technically understood – they just need funded effort to implement known solutions



Science and
Technology
Facilities Council

Questions

Facebook: Science and
Technology Facilities Council

Twitter: @STFC_matters

YouTube: Science and
Technology Facilities Council