New sequencing Storage

Guy Coates

Wellcome Trust Sanger Institute

gmpc@sanger.ac.uk

wellcome trust
sanger
institute

About the Institute

Funded by Wellcome Trust.

- 2nd largest research charity in the world.
- ~700 employees.

Large scale genomic research.

- Sequenced 1/3 of the human genome (largest single contributor).
- We have active cancer, malaria, pathogen and genomic variation studies.

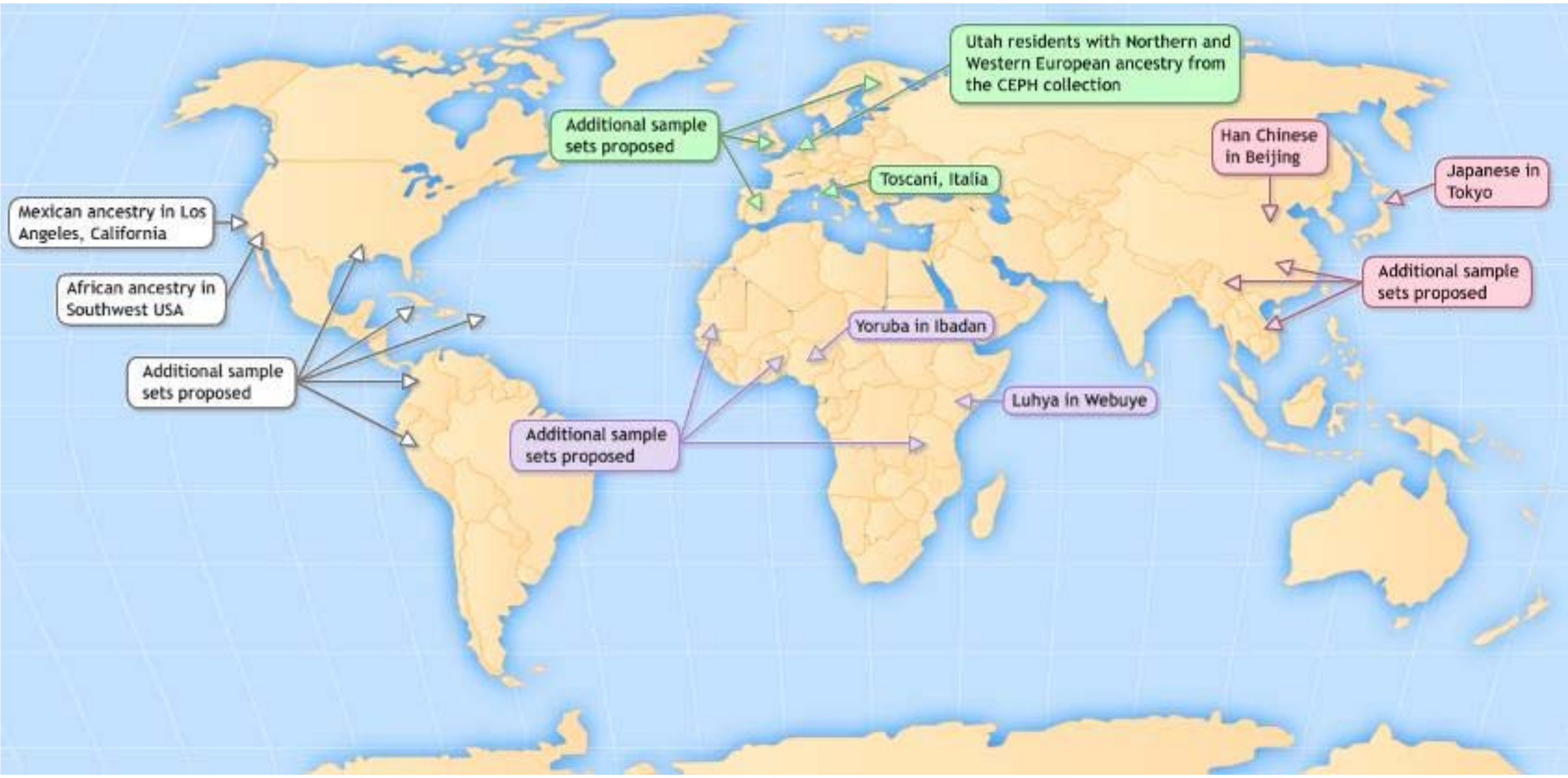
All data is made publicly available.

- Websites, ftp, direct database. access, programmatic APIs.

wellcome trust sanger institute

Recent initiatives: 1

1000 Genomes


Utah residents with Northern and Western European ancestry from the CEPH collection
Additional sample sets proposed
Han Chinese in Beijing
Toscani, Italia
Japanese in Tokyo
Mexican ancestry in Los Angeles, California
Additional sample sets proposed
African ancestry in Southwest USA
Yoruba in Ibadan
Additional sample sets proposed
Luhya in Webuye
Additional sample sets proposed

Overview

1. Scramble for nex-gen sequencing
2. The data explosion
3. Building flexible systems
4. Future directions

Scramble for Next-gen sequencing

Classic Sanger "Stealth project"

Summer 2007; first early access sequencer.

Not long after:

- "15 sequencers have been ordered. They are arriving in 8 weeks. Can we have some storage and computers?"

A fun summer was had by all!

What are we dealing with?

It all started getting very Rumsfeld-ian:

“There are known knowns.”

- Must be in place by Oct 1st.
- ~5TB per week per sequencer.

“There are known unknowns.”

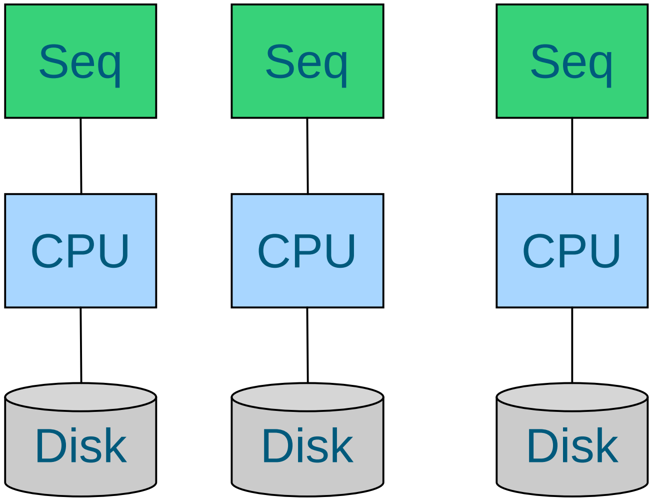
- How does the analysis work?
- Will we be CPU bound or IO bound?
- What will the growth rate be?

“But there are also unknown unknowns.”

Modular

Small: 1 machine + disk pool per sequencer.

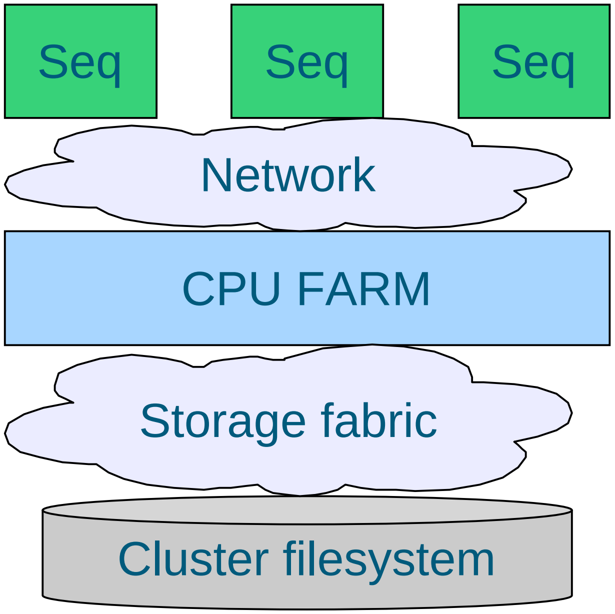
- Simple.
- Easy to scale.
- Small unit of failure; system failure should only affect one sequencer.
- Hard to right-size, especially if things change.



Monolithic

Large pool of machine and storage for all sequencers.

- Flexible: Can cope with fluctuations in CPU/storage requirements.
- Complicated. Clustered storage at scale is hard.
- Eliminating SPOF is hard.
- Scale out can be expensive.



Supersize me

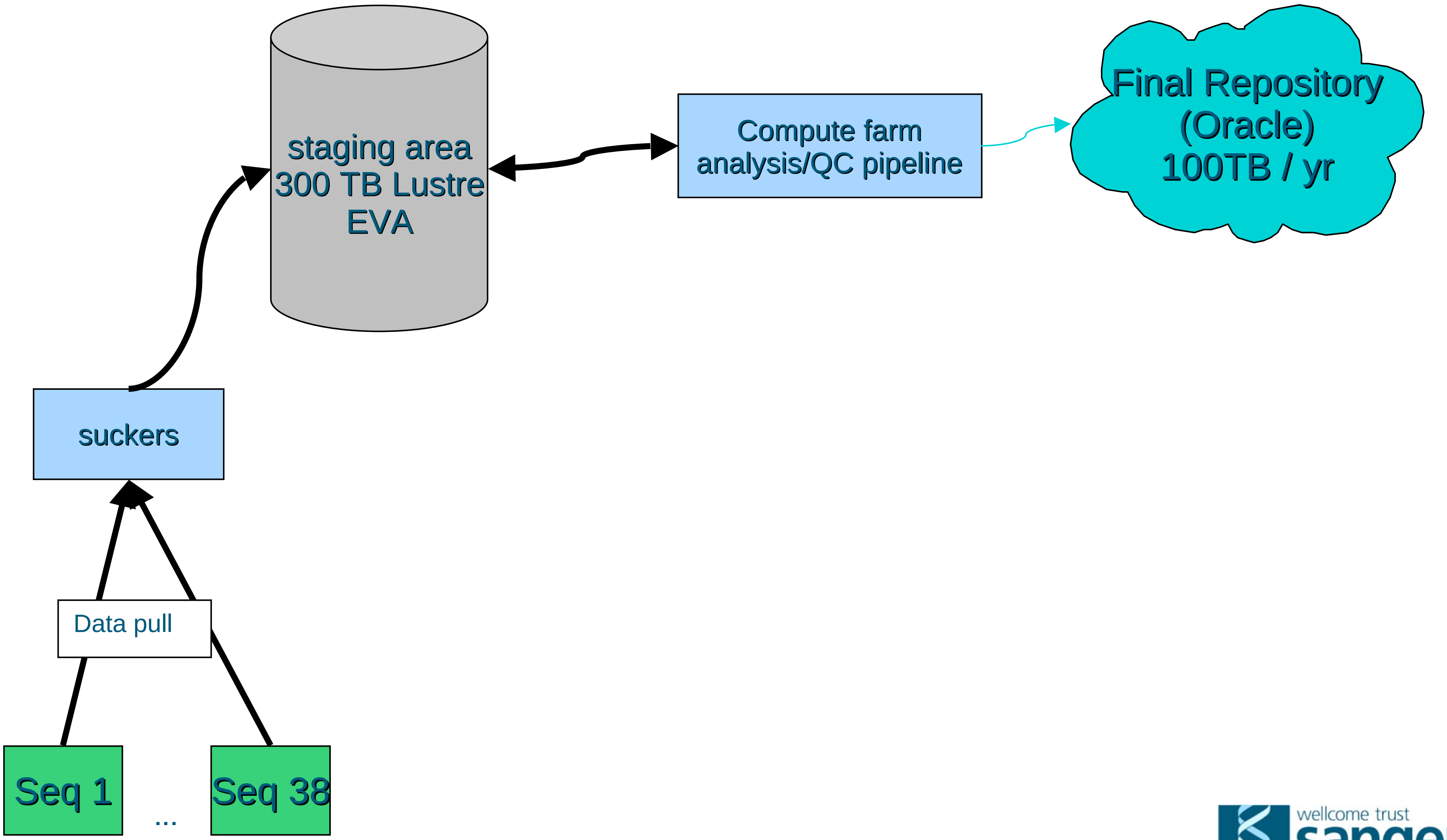
We went monolithic.

- We had experience with large storage and large clusters on our compute farm.

Sequencing Compute farm.

- 512 cores for pipeline and downstream analysis.
- 384 cores for the sequencers, the rest for other analysis.

- 300TB of storage.
  - 3x100TB lustre file-systems (HP SFS / lustre 1.4)
  - Lots of performance if we need it.
  - Plenty of space to cope with changing needs.
- Lustre storage was scratch space.
  - Intermediate image data thrown away after 1 month.
- Final data (sequence+quality) kept in an oracle data warehouse.

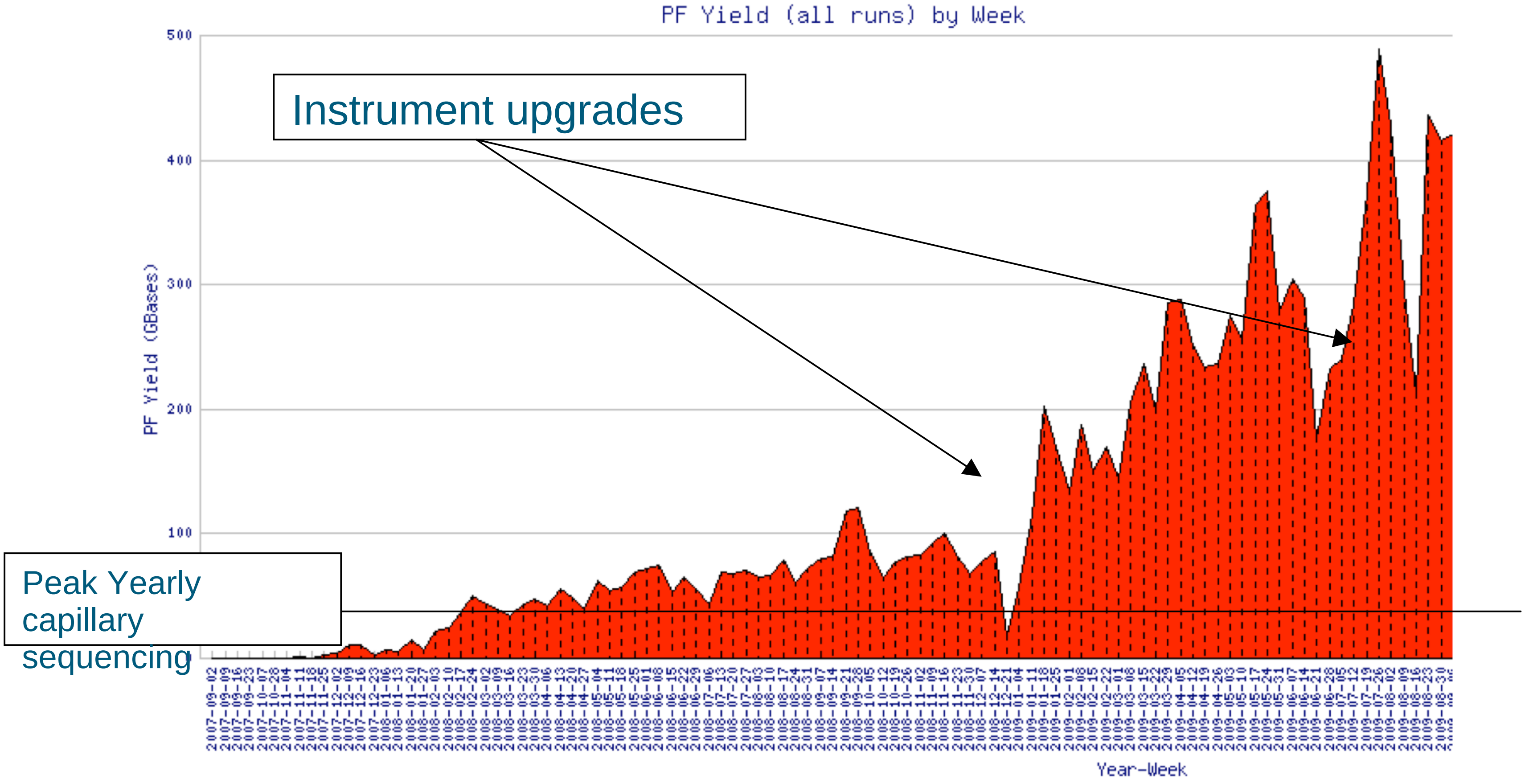

staging area
300 TB Lustre
EVA
Compute farm
analysis/QC pipeline
Final Repository
(Oracle)
100TB / yr
suckers
Data pull
Seq 1
...
Seq 38

Hurrah!

Sequencing storage problems solved, back home in time for tea and medals.

The Data Explosion

The scary graph



Instrument upgrades

Peak Yearly capillary sequencing

Sequencing is not everything...

We had been focused on the sequencing pipeline.

- Taking instrument output and producing DNA sequence + quality.

For many investigators, finished sequence is where they *start*.

- Investigators take the mass of finished sequence data and start computing on it.

Big increase in data all across the institute.

© Barcroft Media

Expanding Everything

Sequencers:

- 15 → 38, numerous upgrades, run-length increases, paired end etc..

Sequencing compute farm:

- ~500 → 1000 cores
- 300 → 600TB
  - Data retention time 4 weeks → 2 weeks.

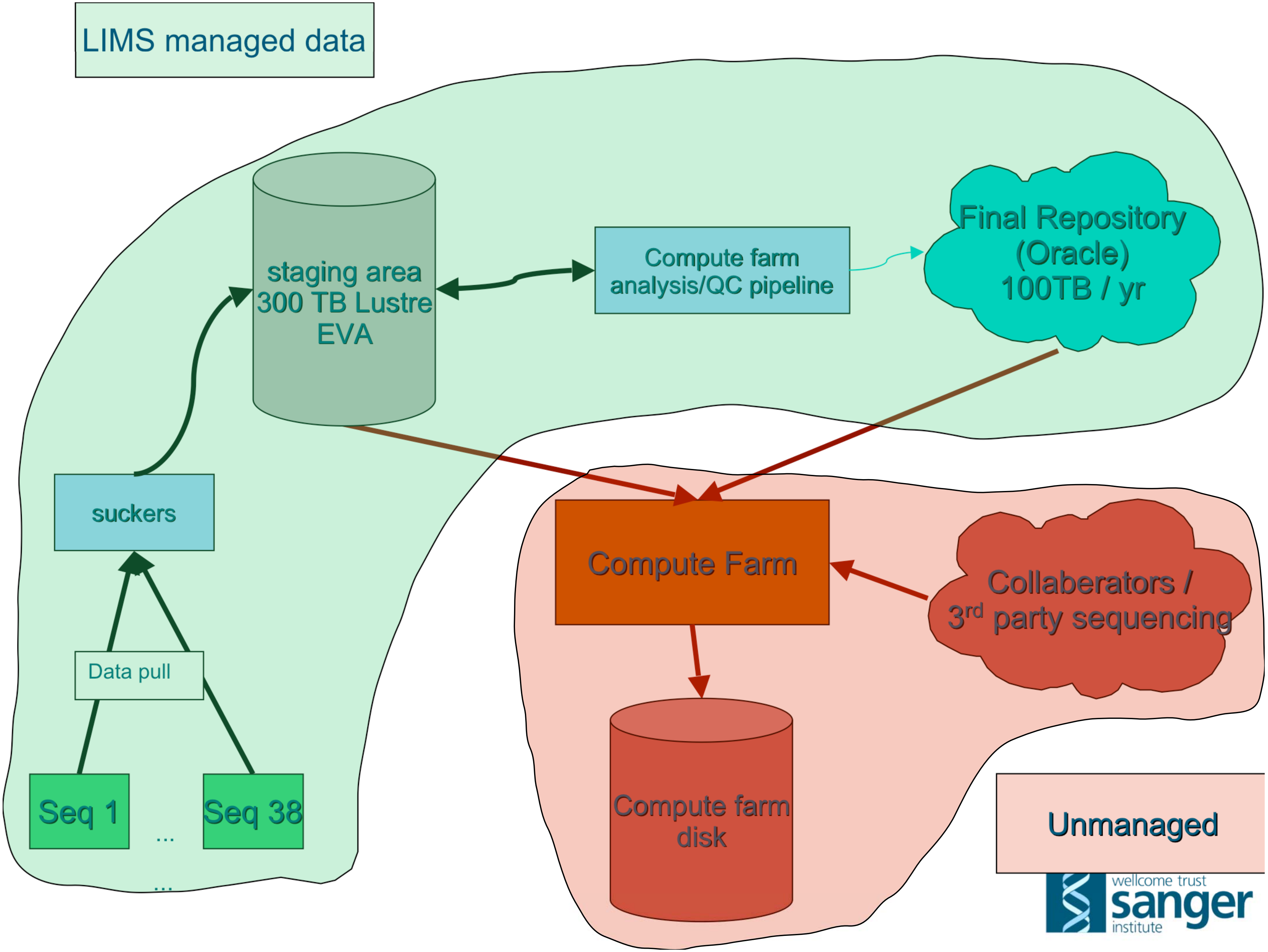
General compute farm:

- 2000 → 5000 cores
- 20TB → 300 TB

General storage

- 2PB → 4PB

Most of the increases were not in the "Sequencing" area.



LIMS managed data

staging area
300 TB Lustre
EVA

Compute farm
analysis/QC pipeline

Final Repository
(Oracle)
100TB / yr

suckers

Data pull

Seq 1

...

Seq 38

...

Compute Farm

Collaberators /
3rd party sequencing

Compute farm
disk

Unmanaged

wellcome trust
sanger
institute

Pipeline data is managed

Data in the sequencing pipeline is tracked.

- We know how much there is, and who it belongs to.

Data has a defined life-cycle.

- Intermediate data (images etc) are deleted after the runs pass QC.
- Important data (finished sequence) is automatically moved to our archive, backed up and replicated off-site.

Good communication between the pipeline / LIMS developers and the systems team.

- We know who to talk to.
- We get a good heads up for changes/plans.

Unmanaged data is bad...

Investigators take data and “do stuff” with it.

- Analysis take lots of space; 10x the space of the “raw” data.

Data is left in the wrong place.

- Typically where it was created.
  - Moving data is hard and slow.
- Important data left in scratch areas, or high IO analysis being run against slow storage.

Capacity planning becomes impossible.

- Who is using our disk space?
  - “du” on 4PB is not going to work...
- Are we getting duplication of datasets?

How do we account for it?

- We need to help Investigators come up with costings that include analysis costs as well as the costs for initial sequencing.

wellcome trust sanger institute

Old architecture

Separate compute silos for separate groups

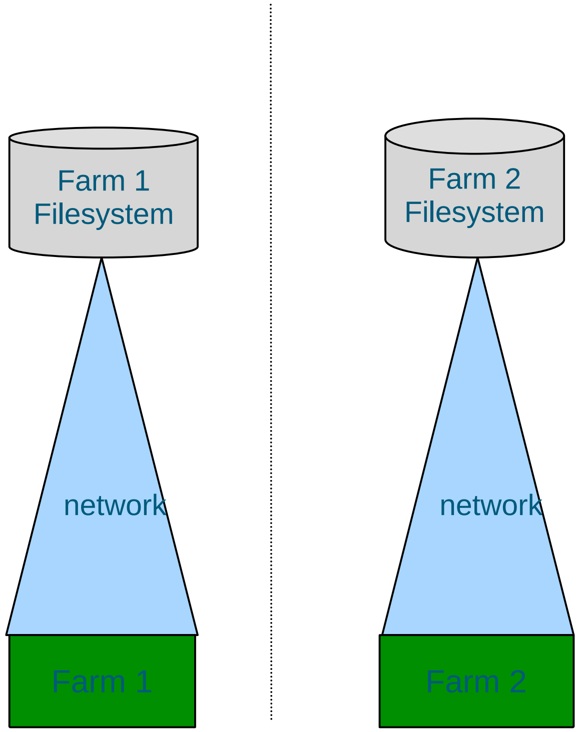
- Eg cancer, pathogen, sequencing

“Fast” cluster storage (lustre) for IO bound work.

- Constrained by network topology/bandwidth.

Problems:

- Projects span the organisation domains.
- Projects can outgrow compute + storage.
- Fast storage is always full.
- Cannot afford to buy more.



Another Scary Graph

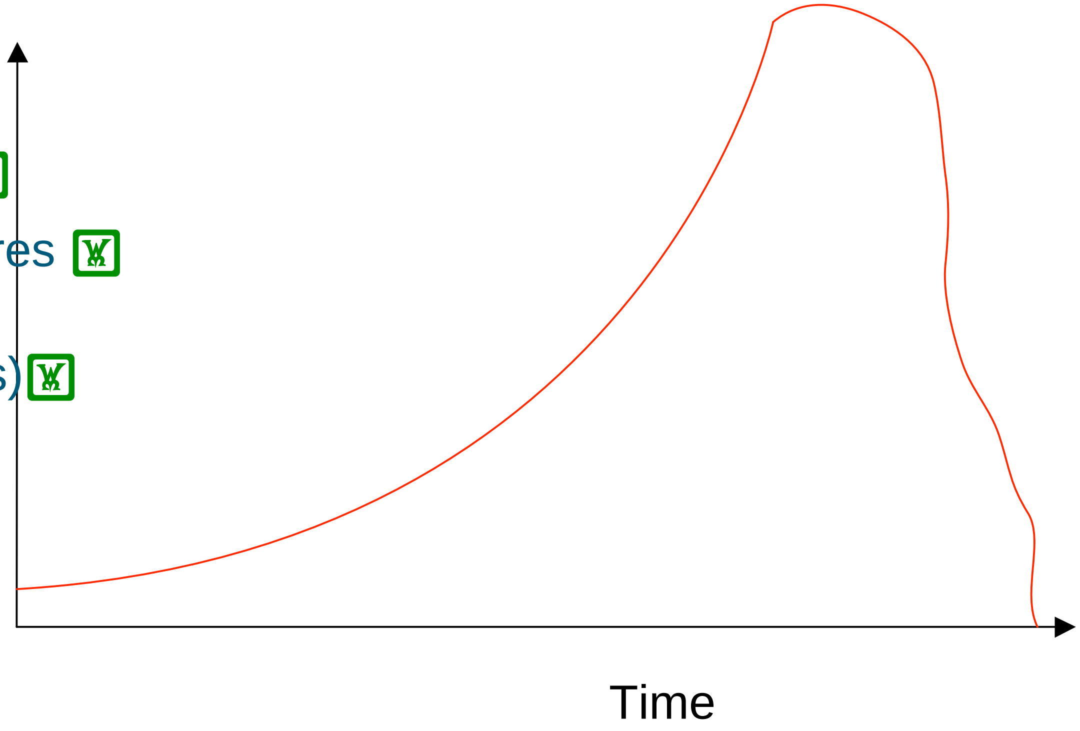
Storage 4PB

Budget

Compute 5K cores

Headcount

Sequencing output (xTBases)



Building Flexible Systems

wellcome trust
sanger
institute

Agile

Science is changing very rapidly.

- Changing science usually means more data.

LIMS / Pipeline software development teams use agile methods to cope with changes.

- Short, iterative, development process.
- bi-weekly updates to pipeline.
- Evolutionary process.
- Allows changes to be put in place very quickly.

Can we do “agile” systems?

- If systems can adapt easily, we do not have to worry about changes in science.

Buzzword compliant!

Plan of attack

1. Look at the workflow and data
2. Start managing data
3. Tie it all together into a flexible infrastructure

Workflows

Identified 3 data patterns:

High IO, active datasets.

- Data being crunched on our compute farm.
- Needs high performance storage.

“Active Archive” datasets.

- Projects that has been finished, but are still need to be around.
  - Reference datasets, previous tranches of data.
- Does not need to be fast, but it needs to be cheap, as we are going to have **lots** of it.

Stuff in the middle.

- Home directories etc.

High speed disk

Lots of options around for high speed cluster storage.

- Lustre, GPFS, Isilon, Panasas, pNFS etc.
- These are exotic and/or expensive.
- Expect to break them and expect to spend time on care and feeding.
- But you need them at scale.
- We use DDN + Lustre 1.6 + support contract.

Single name-space file-systems across clusters are nice; our investigators really like them.

- when they work:)

Low speed / bulk disk

This will be the bulk of our data; we needs lots of it.

- Price / TB is critical factor.
- Shortly followed by space and power footprints.
- Power and space are constraints for us.
- Dense disk shelves.
- MAID functionality.
- Needs to be manageable.

And don't forget the backup.

- Backup to tape is probably not practical.
- Use disk → disk replication.
- You need 2x of whatever you buy.
  - Do we do this in hardware? Software? Both?

wellcome trust
sanger
institute

Start Managing data

How do we distribute data between these two disk pools?

Manually.

- This is less than idea, but you have to start somewhere.
- Work with the researchers to identify data and then map it onto storage.
- Works well with the power users.
- Can be difficult with transient projects who do one-off large-scale analysis.
  - Data is orphaned.
  - It takes along time to track down who is actually responsible for the data.
- Stick rather than carrot; quotas, limits.

Still a massive improvement.

Flexible Infrastructure

Make storage visible from everywhere.

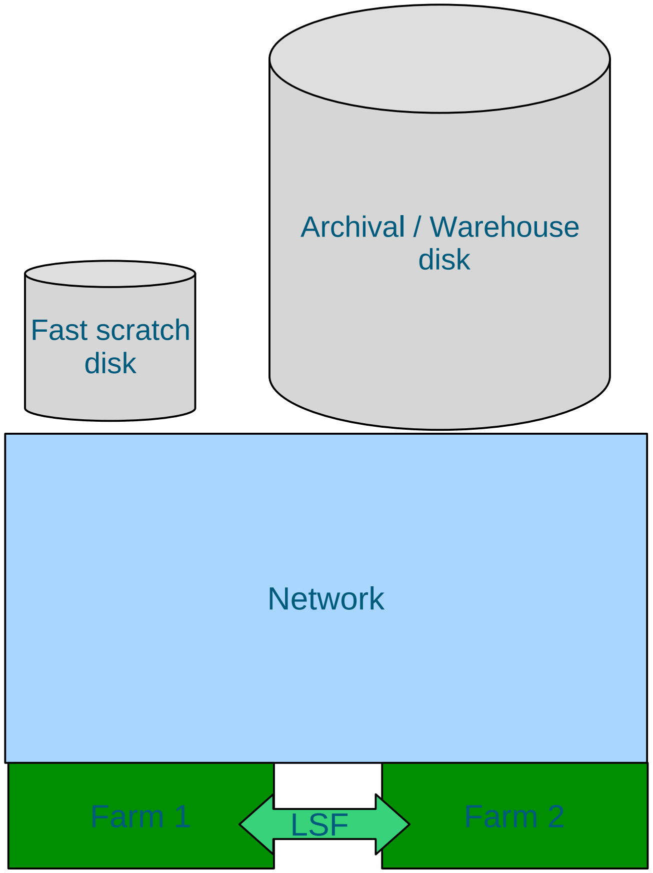
- Key enabler; lots of 10Gig

This allows us to move compute jobs between farms.

- Logically rather than physically separated.
- Currently using LSF to manage workflow. VMs in the future?

Modular design.

- Blocks of network, compute and storage.
- Assume from day 1 we will be adding more.
- Expand simply by adding more blocks.



Future Directions

A.K.A.: Stuff that looks good in powerpoint but we still have to actually do.

Modular Systems

Modular approach for compute and network in place.

- Add racks / chassis of compute and tie it all together with a load of 10GigE networking.
- OS management / deployment tools take care of managing configs for us.

Fast disk modules in place.

- DDN / lustre.

We have not identified our building block for the bulk storage.

- How big should a storage module be?
  - Storage fails (even enterprise storage).
  - What is the largest amount of storage we are comfortable with losing?
- How much disk behind a controller?
  - Effects price / TB dramatically.
- And all the other features.
  - MAID, dense, power efficient, scalable, replicable, etc

Data mangement

Data management / movement needs to be done automatically.

- People are error-prone, software is not.
- We need some sort of HSM / ILM systems.

Can we find one that:

- Works at scale: (10s of PBs, Billions of files)
- Does not tie us in to a file-system/storage setup that will be obsolete in 5 years.

How far should we empower end users to manage their data?

- They know the most about their data and their workflow.
- However, they are scientists, not sys-admins.
- Their data is our responsibility.

Cloud

Cloud...

Cloud as compute on demand:

- We have spiky compute demands.
  - Especially when “stealth” sequencing or analysis projects break cover.

Cloud as an external datastore:

- Cheap, off-site data archiving. (Disaster recovery).

Cloud as a distributed datastore:

- Our data is publically available.
  - Buy downloading 5TB of data across the public internet is not a pleasant experience.
- Cloud providers can do worldwide replication/ content delivery.

Cloud as a collaboration:

- Data on its own in not very useful.
- Bundle data and analysis pipelines for others to use.

Our Cloud Experiments

Can we take an Illumina run (4TB), run the image analysis pipeline and then align the results?

Pipeline ran, but it needed some re-writing.

- IO in Amazon is slow, unless you use S3.
  - NFS on EBS performance is unusable with > 8 clients.
- S3 is not POSIX.
  - Even with FUSE layer, code re-writing required.

Cloud compute is easy, cloud storage is hard.

Getting data in and out is slow:

- We realised ~10% of our theoretical bandwidth to Amazon.
  - Even with gridFTP.

Promising, but more work needed...

Acknowledgements

Sanger System

- Phil Butcher

Informatics Systems Team

- Pete Clapham
- James Beal
- Gen-tao Chiang

wellcome trust
sanger
institute