
2012 Bio-IT World Webinar with Aspera

Chris Dwan cdwan@bioteam.net

Bioteam (<http://bioteam.net>)



The BioTeam Inc.

Independent Consulting Shop

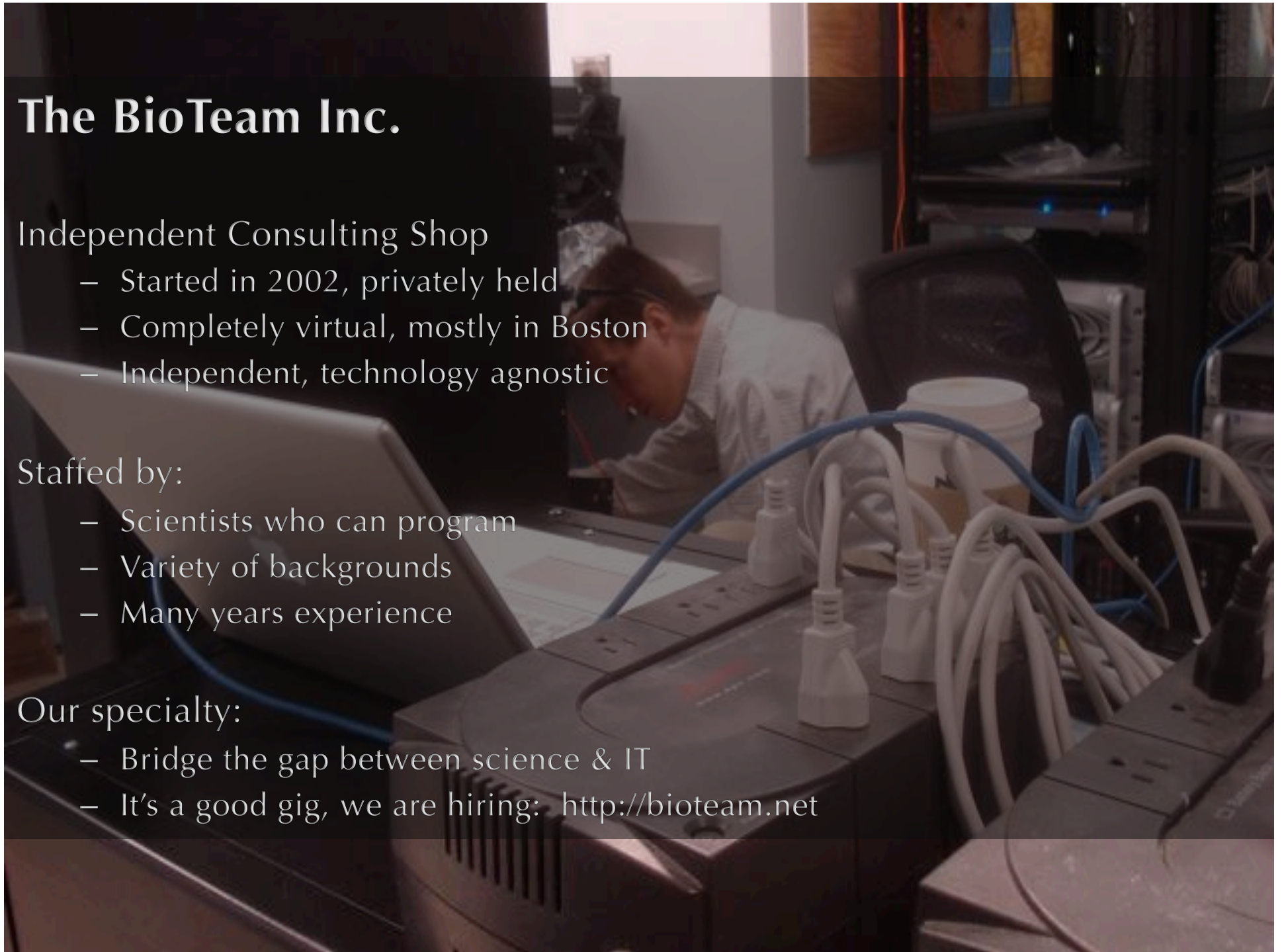
- Started in 2002, privately held
- Completely virtual, mostly in Boston
- Independent, technology agnostic

Staffed by:

- Scientists who can program
- Variety of backgrounds
- Many years experience

Our specialty:

- Bridge the gap between science & IT
- It's a good gig, we are hiring: <http://bioteam.net>



New York Genome Center

- Collaborative effort between 12 New York institutions
 - \$1 × 10⁸ capital raised from academic, corporate and philanthropic sources
 - Novel collaboration between Clinical, Academic, Pharma research
- Potentially the largest facility of its kind in North America
 - Initial footprint of 36+ HiSeq 2000 instruments, scaling to 100+ by 2017
 - 500+ in-house staff, heavily weighted to bioinformatics
 - Initial service offerings in spring 2012
 - Manhattan facility open in late 2012
- It's a good gig, we are hiring: <http://nygenome.org>

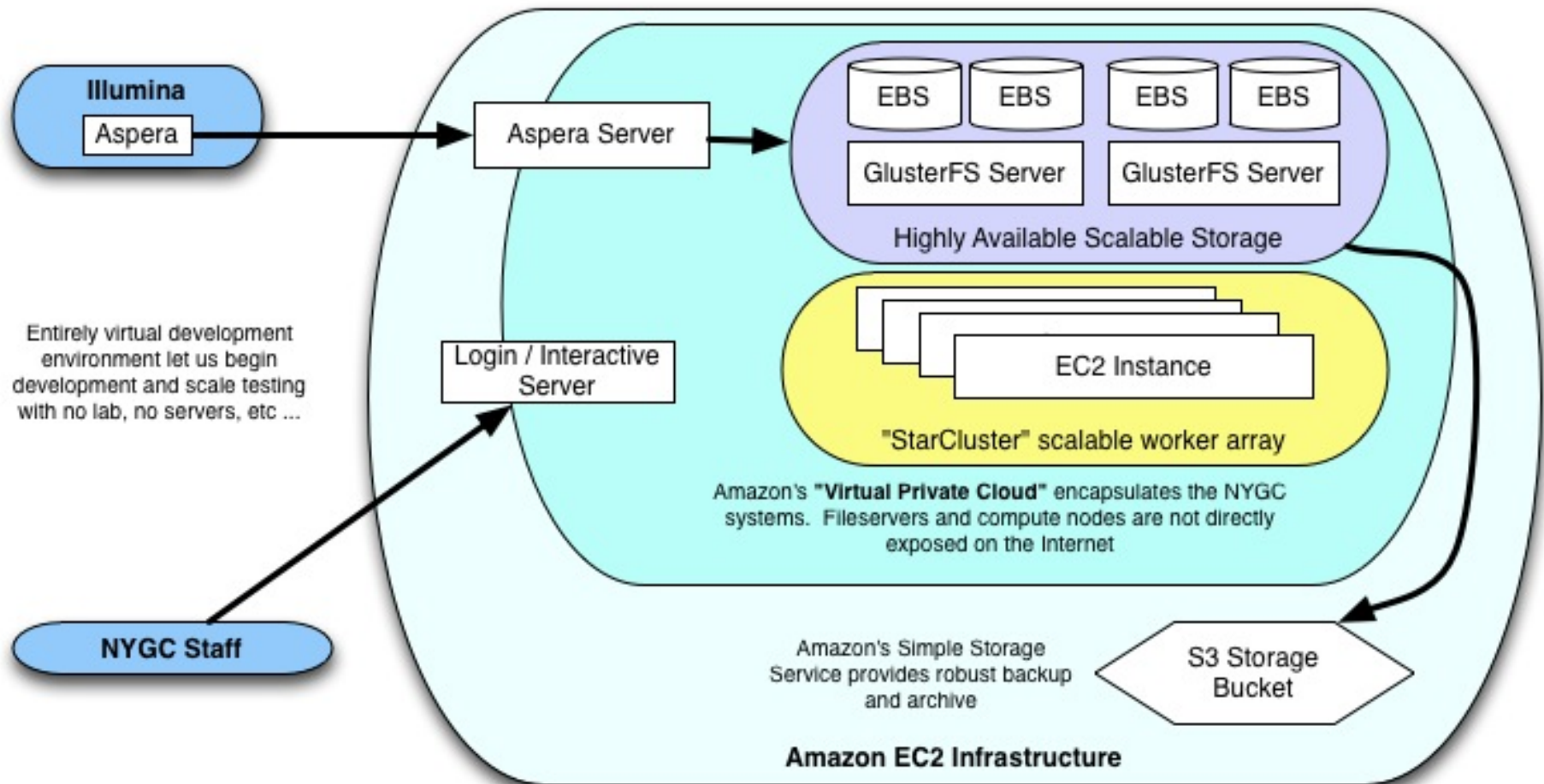
Genome Scale Data: The Sky is NOT falling

- ~130GB per “human genome” in raw BAM format
 - 3×10^9 base pairs in a human genome
 - 30x coverage (short reads)
 - 1.25 bytes per base pair (BAM file, including quality information)
 - Nobody stores raw TIFF images anymore
 - Emerging technologies will achieve massive reductions (CRAM)
- NYGC:
 - Several batches up to 1,000 genomes this year
 - Exome and RNA-SEQ are “genome lite” for data volumes
 - Sequencer capacity scaling to ~130 HiSeq 2000's (7×10^{12} Tb/day) by 2017
 - Data archive scaling to 5 – 10PB over the next several years

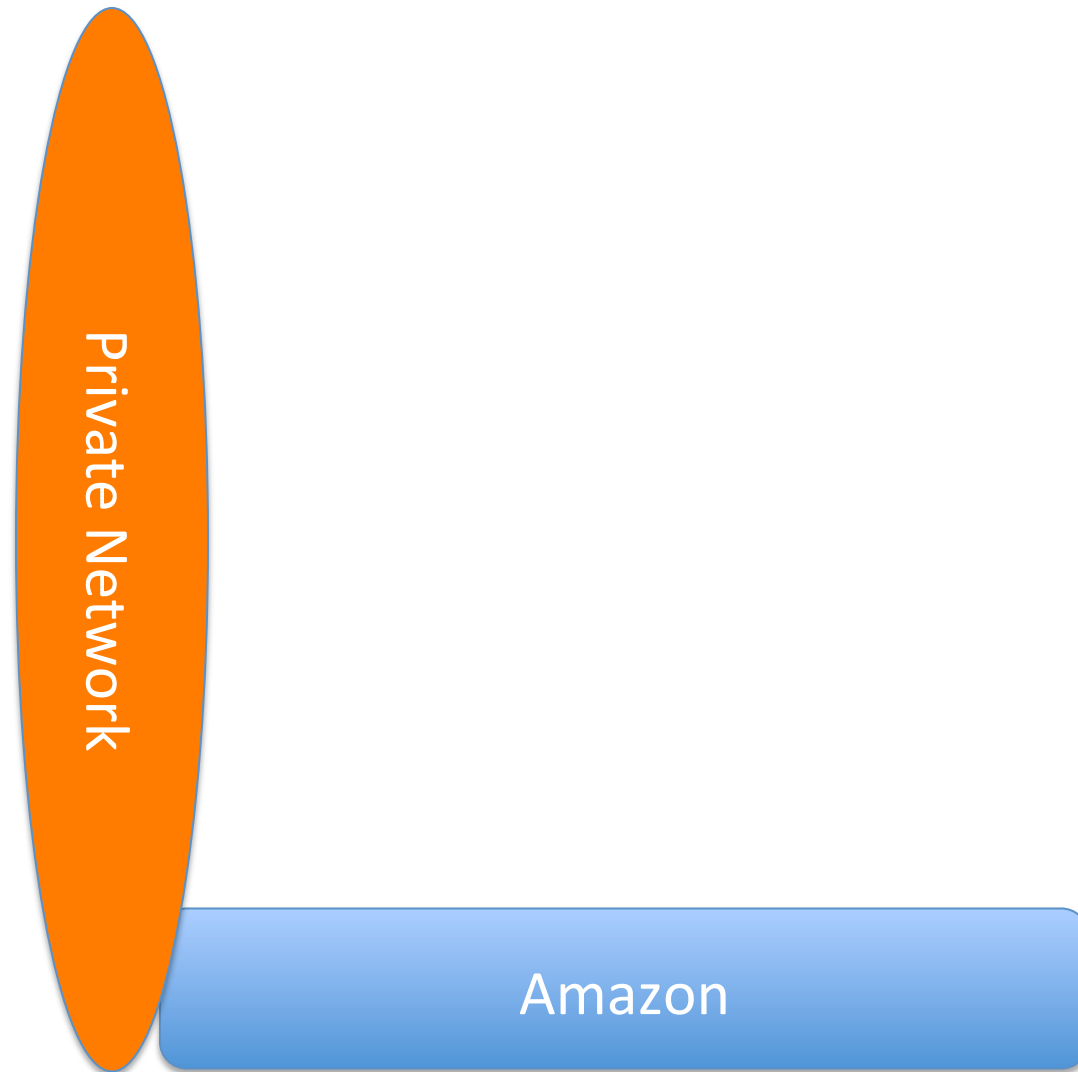
2012: Petabyte scale storage is engineering, not research

First samples: No physical infrastructure

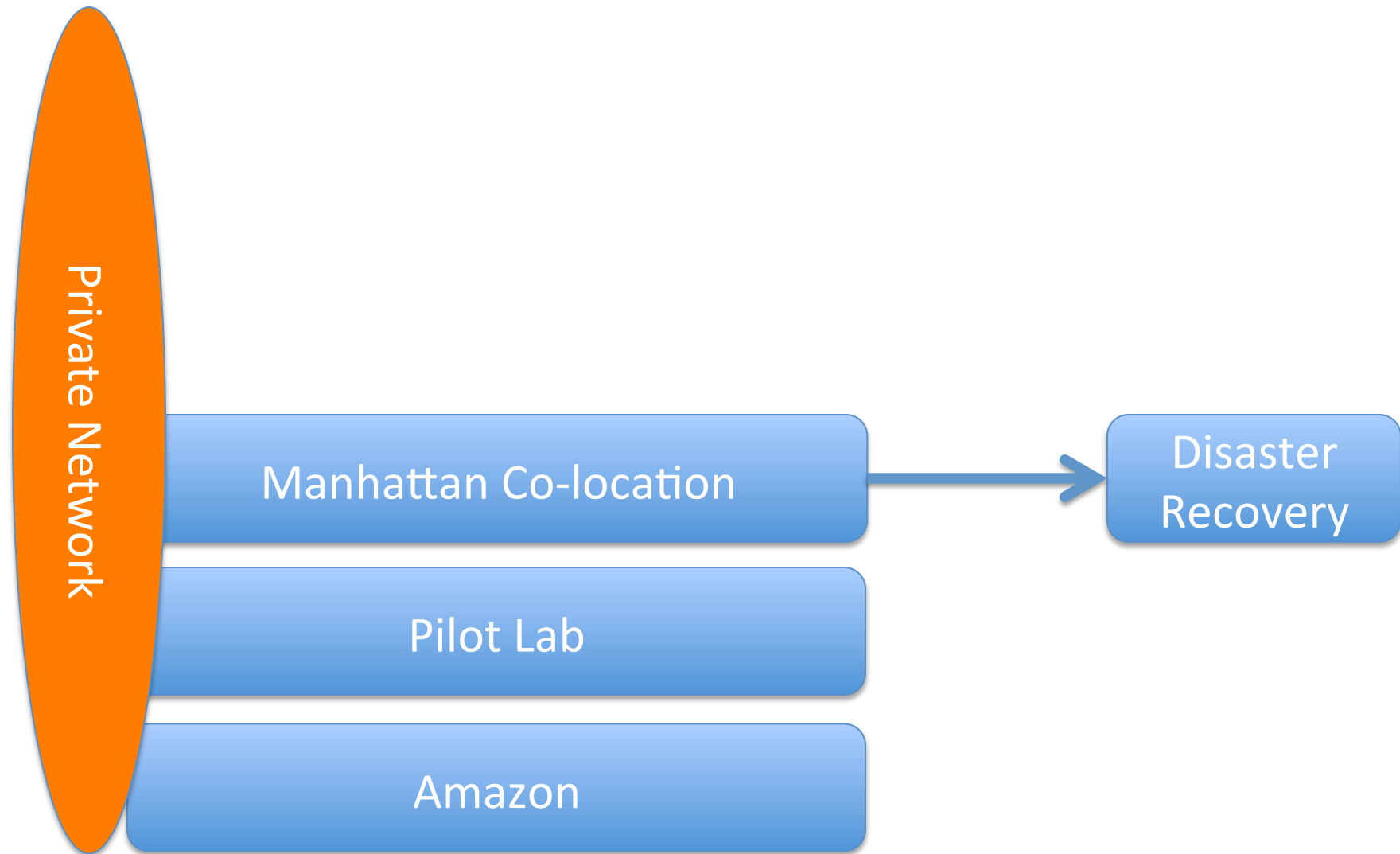
Virtual infrastructure for NYGC, prior to physical lab space or customers



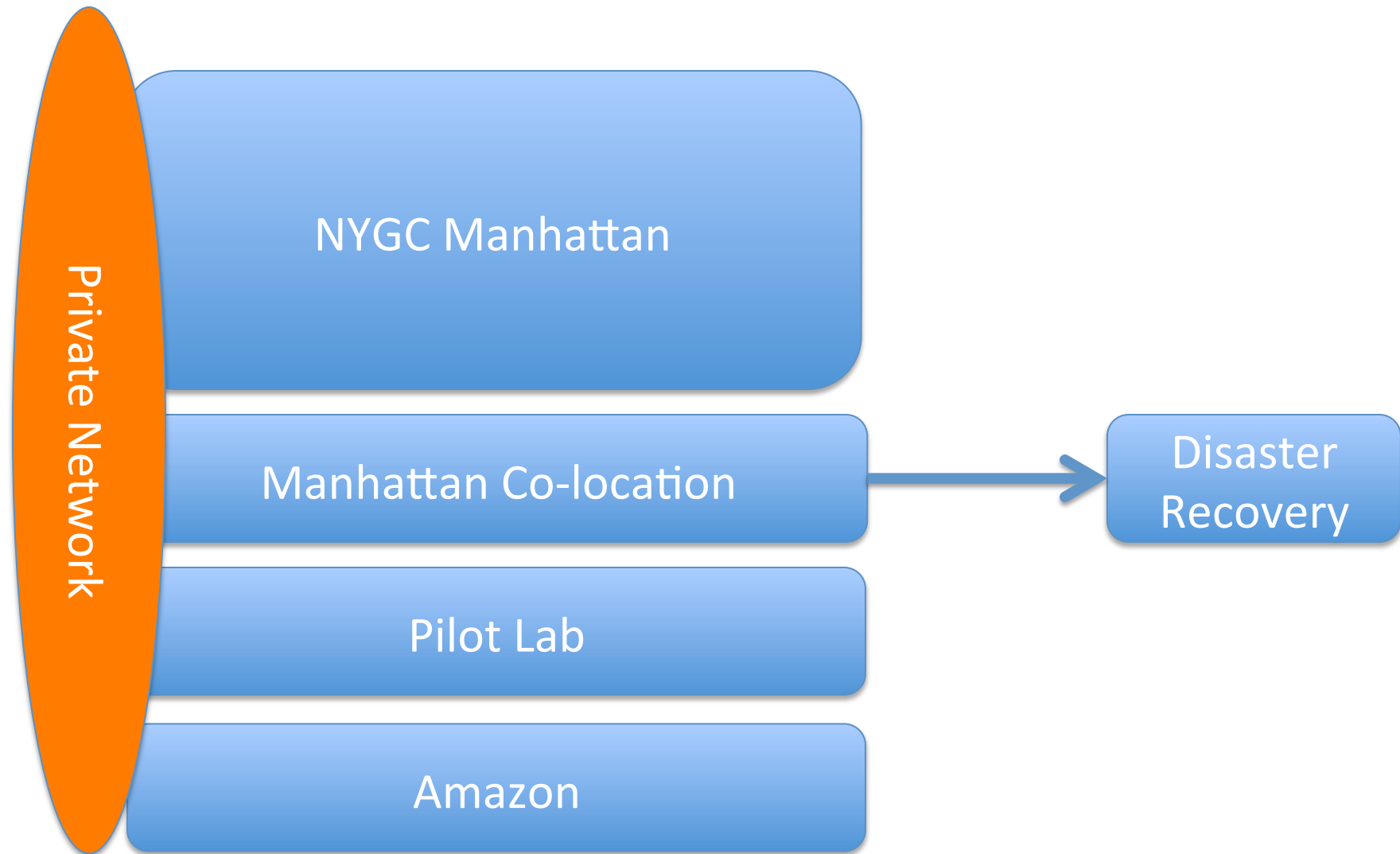
Infrastructure-bursting



Infrastructure-bursting



Infrastructure-bursting



Objections

- The “public cloud” is unproven and unsafe.
 - Also, CLIA, HIPPA, etc.
 - Wouldn't you prefer a nice “private cloud” instead?
- The data are too large, we must ship disks via FedEx.

Amazon Infrastructure as a Service (IaaS)

- Stop saying “cloud” like it means something.
 - “We’ll use science!”
- Amazon’s S3 data storage is more robust and reliable than what you can build.
 - 99.999999999% durability
 - 762×10^9 objects stored at the end of 2011
 - 500,000 requests per second
- HIPPA / CLIA certifications are possible
 - Don’t let the Fear Uncertainty and Doubt crowd tell you otherwise.
- It is still engineering, but high levels of security and operational availability are possible.

Shipping disks via Fedex

- Assume a 48 hour point to point latency
 - Two sets of checksums
 - Sneakerbot / human interaction on either side
- Very low potential for automation
- Many opportunities for error
- Sadly, this is the state of the art in many places



Data Bandwidth

Bandwidth	1 Gigabyte	1 Genome (130GB)	Genomes / day	HiSeq 2000 daily raw output (55GB/day)
T1 business link (12Mb/sec)	11m 22s	24.6h	1	2
T3 business link (45Mb/sec)	3m 10s	6.9h	3.4	9
700Mb/sec	11s	24m	60	134
Gigabit	8 sec	17m	84	192



2017 capacity of the center,
in terms of 2012 instruments

If we can make full use of the available bandwidth, gigabit networking is sufficient for the long term data motion needs of the center



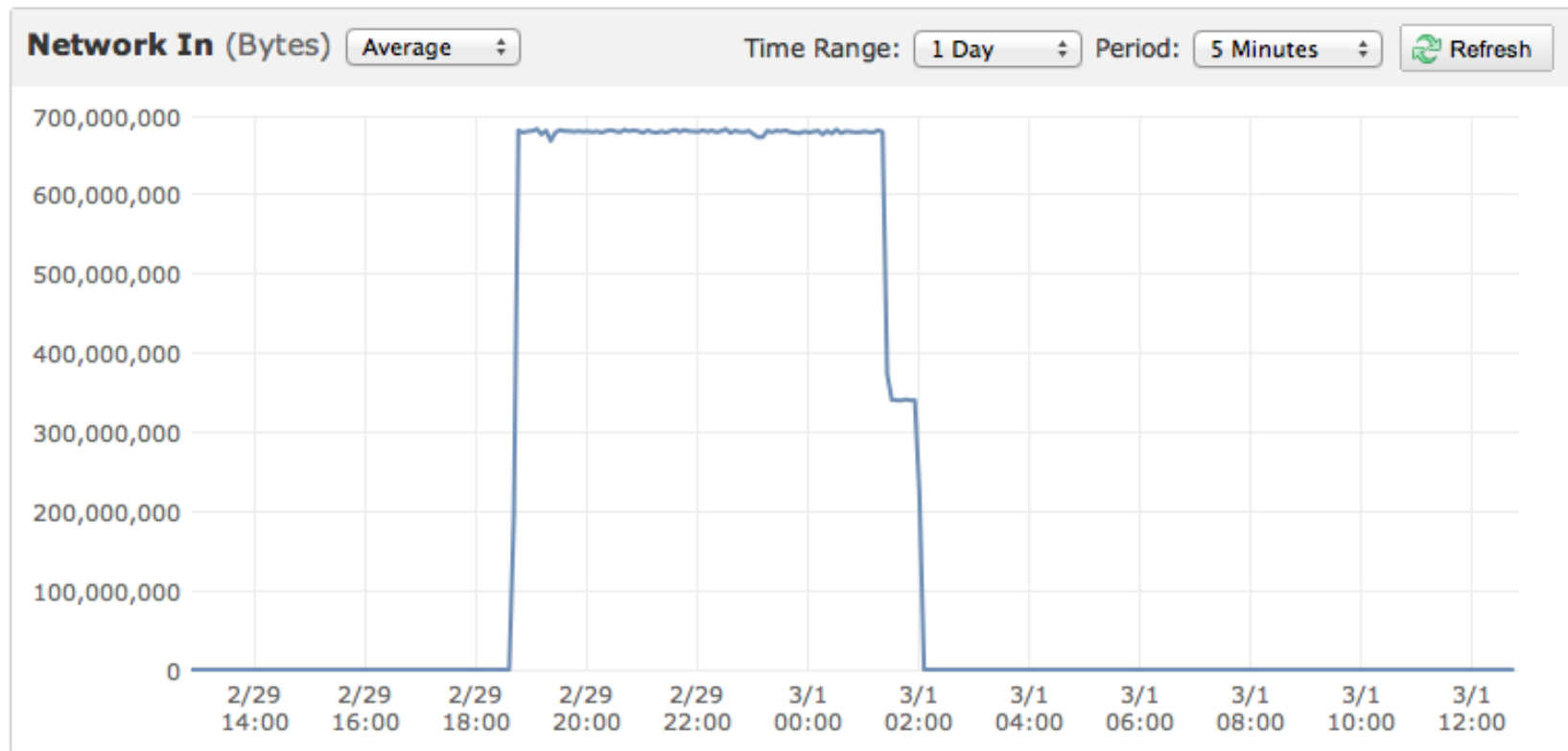
Aspera “*FASP*”

- TCP based data motion is fundamentally limited by packet loss and latency
 - FTP and HTTP always seem to self limit around 20MB/sec
- Eventually, everybody writes a parallel FTP / rsync / scp wrapper
 - Avoiding the problem, not solving it
 - Assumptions of independent errors and paths
 - Yes GridFTP, we see you in the back
- Aspera ***FASP*** addresses the problem at the root
 - Scales to 10Gb/sec and beyond
 - Allows fine grained quality of service control / monitoring
 - Early adoption with media customers

My Aspera Deployment Story

- Purchase license via Amazon: \$750/mo, charged under Amazon ID
- Additional \$0.15/GB usage charge (\$20/genome)
- Deploy Server in EC2 from Amazon dashboard (minutes)
- Log in via ssh, configure user account, log in on web, poke around
- Give credentials to collaborator at lab in CA
- **Hiccup 1: Rate limited at 5 Mb/sec**
 - Replace faulty router at collaborator site
- **Hiccup 2: Rate limited at 45Mb/sec**
 - Default in Aspera software set to 45Mb/sec.
- **700Mb/sec on the third try with no real tuning**

Aspera Data Motion (coast to coast)



Monitored Instances: ■ i-772cfe12


Times are displayed in UTC.

Close

Note: datapoints are plotted at the start of the period.

Data Bandwidth

Bandwidth	1 Gigabyte	1 Genome (130GB)	Genomes / day	HiSeq 2000 daily raw output (55GB/day)
T1 business link (12Mb/sec)	11m 22s	24.6h	1	2
T3 business link (45Mb/sec)	3m 10s	6.9h	3.4	9
700Mb/sec	11s	24m	60	134
Gigabit	8 sec	17m	84	192



Observed coast to coast over the commercial internet



2017 capacity of the center, in terms of 2012 instruments

Downstream

- “Direct to S3” storage
 - Object storage that looks like a filesystem to the user
- “Shares”
 - Integrated browsing / access over multiple data repositories (both AWS and data center)
 - Persistent URI to access data, no matter where it has moved
 - Integration with one or more AD / LDAP authentication servers
- ...



end;