Gigabyte Bandwidth Enables Global Co-Laboratories

Prof. Harvey Newman, Caltech
Jim Gray, Microsoft

Presented at Windows Hardware Engineering Conference
Seattle, WA, 2 May 2004

Microsoft Windows Hardware & Driver Central

AMD

Credits:
This represents the joint work of many people.

CERN
- Olivier Martin
- Paolo Moroni.

Caltech
- Julian Bunn
- Harvey Newman
- Dan Nae
- Sylvain Ravot
- Xun Su
- Yang Xia;

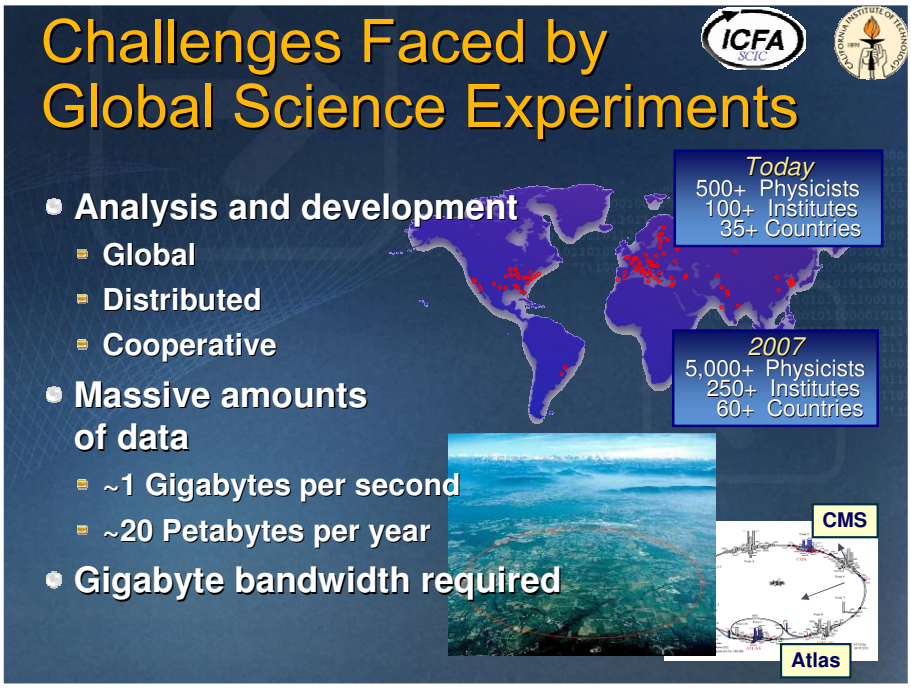
S2io
- Leonid Grossman

AMD
- Brent Kelley

Newisys
- John Jenne
- Doug Norton
- Rich Oehler
- Dave Raddatz

Microsoft
- Eric Beardsley
- Jim Gray
- Neel Jain
- Peter Kukol
- Clyde Rodriguez
- Inder Sethi
- Ahmed Talat
- Brad Waters
- Bruce Worthington

Ordinal
- Charles Koester
- Chris Nyberg



Jim:

Thanks Bill.

Good Morning!

We are honored to have Harvey here today.

He leads the effort connecting the experiments at CERN in Geneva, Switzerland with Physicists in the Americas

and co-leads the International Virtual Data Grid Laboratory.

Harvey, what is the Virtual Data Grid?

Harvey:

Large physics experiments have become international collaborations involving thousands of people.

Current and new experiments share a common theme:

Groups all over the world are distributing, processing and analyzing massive datasets.

The picture at right shows the CERN LHC accelerator, where four experiments will generate about 300 terabytes per second.

Pre-filtering reduces that to about a gigabyte per second of recorded data, or 20 petabytes a year.

Jim:

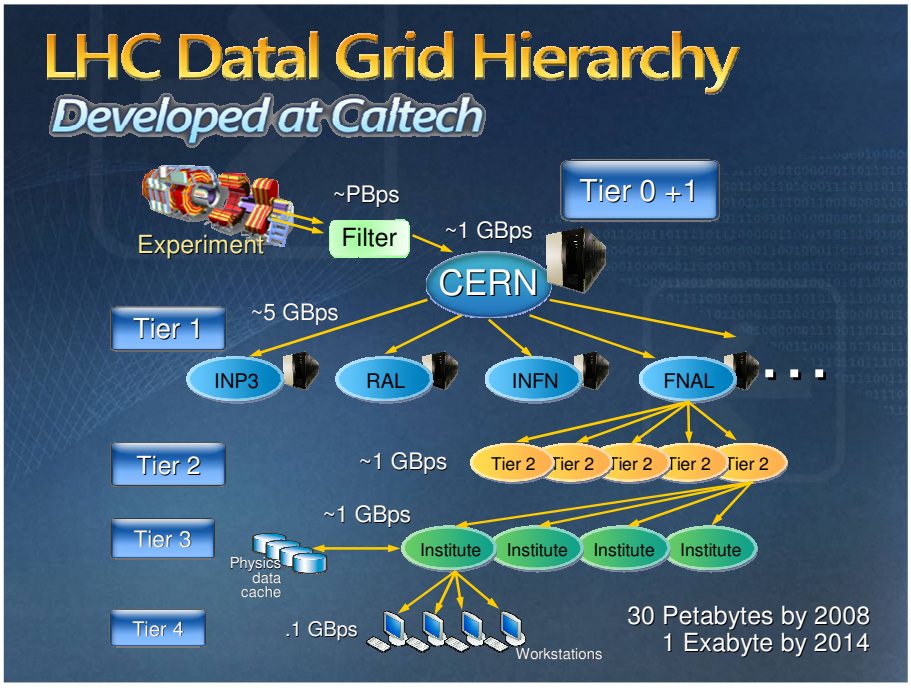
Wow! That's a 2 zetabytes a year.

The equivalent of sifting through a hundred million copies of the Library of Congress each year and saving 1,000 copies of it.

Harvey:

And we continually reprocess and redistribute it, as we advance our ability to separate the rare "discovery signals" from the petabytes of standard physics events we already understand.

That's why we need gigabyte per second bandwidth.



Harvey:

We developed a four tiered architecture to support these collaborations.

Data processed at CERN is distributed among the Tier1 national centers for further analysis.

The Tier1 sites act as a distributed data warehouse for the lower tiers, and refine the data by applying the physicists' latest algorithms and calibrations.

The lower tiers are where most of the analysis gets done.

They are also a massive source of simulated events.

Jim:

As a database guy, I am scared of managing 20 petabytes, that's hundreds of thousands of disks.

Harvey:

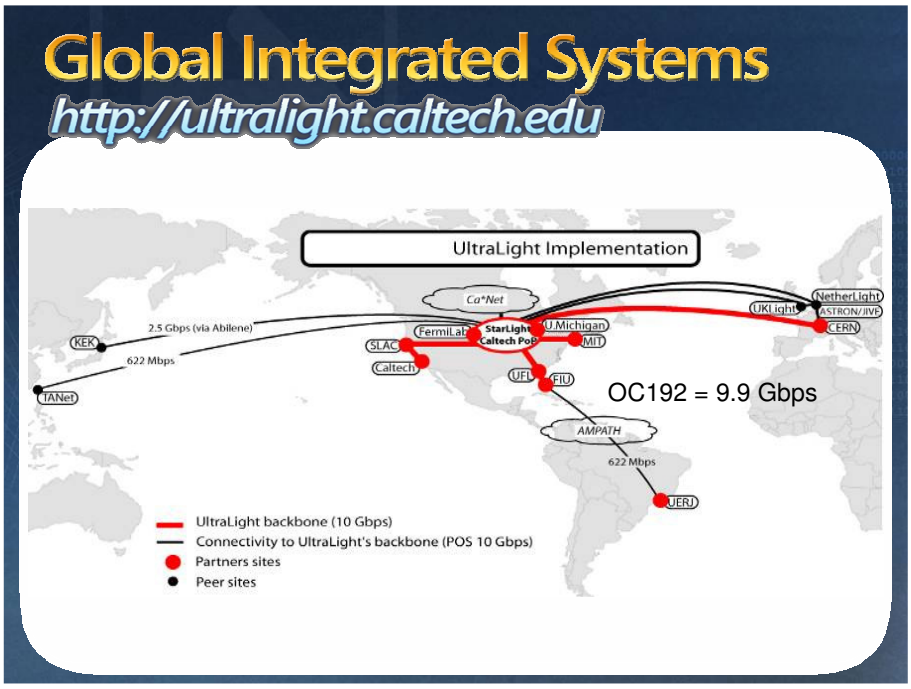
Yes, and as the LHC intensity increases, the accumulation rate will increase, and we expect to reach an Exabyte stored by 2013-15.

All the flows in this picture are designated in gigabytes per second; so it's clear why we need a reliable gigabyte per second network;

And our bandwidth demand is accelerating, along with many other fields of science; growing by a factor of two each year or 1000-fold per decade;

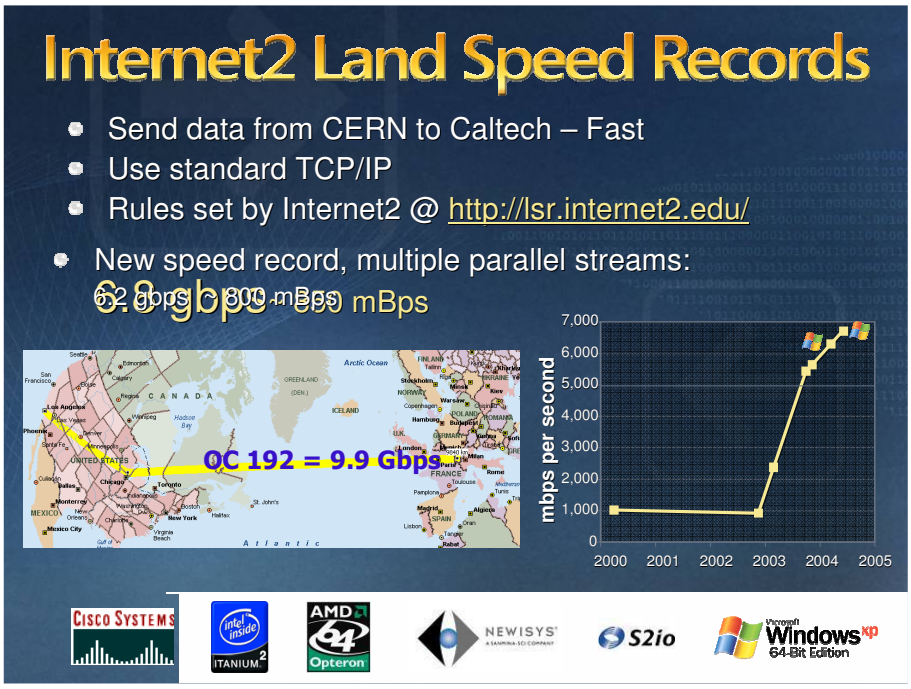
Much faster than Moore's Law. .

We have to innovate each year, learning to use the latest technologies effectively just to keep up.



Harvey

This is the network we're building.
We're working with our partners to extend it round the world.
The next generation will be an intelligent global system,
with real-time monitoring and tracking, and
with adaptive agent-based services at all layers
that control and optimize the system.



Jim:

Internet2 is a consortium of universities building the Next Generation Internet.

They defined a speed contest to recognize work on high-speed networking.

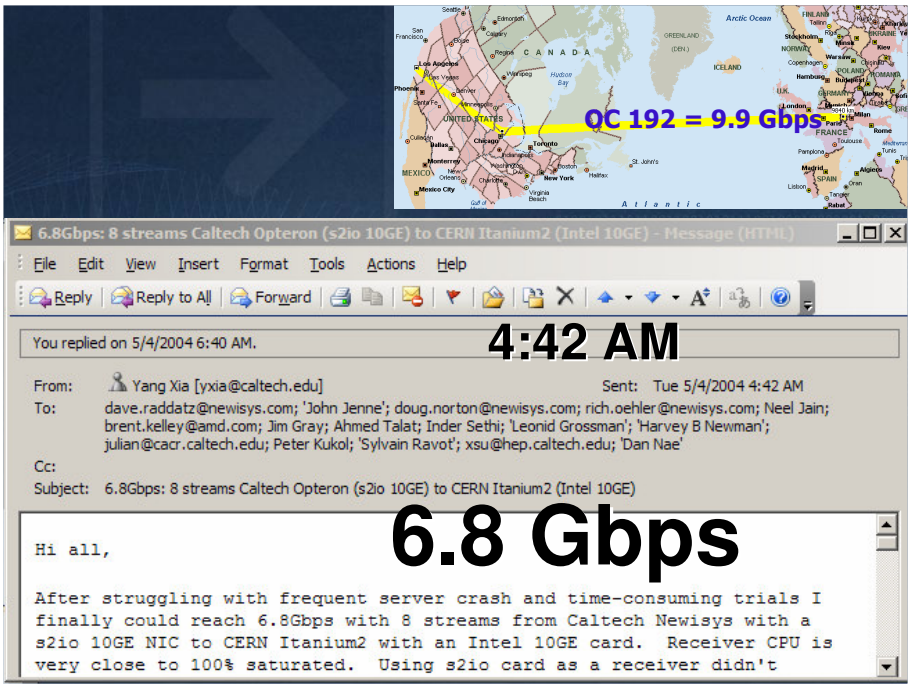
Microsoft entered and won the first round of this contest, but we have not entered since.

Indeed, Harvey's team at Caltech has been setting the records over the last two years.

Harvey:

To set the records, including the one just certified by Internet2, we used the network shown on the previous slide and out-of-the-box tcp/ip to move data from CERN to Pasadena.

Using Windows on Itanium2 servers with Intel and S2io 10 Gigabit Ethernet cards, and Cisco switches, we reached 6.25 gigabits per second – or just under 800 megabytes per second.



Harvey:

Last night we set a new record, actually at 4:30 a.m. using Windows on new Isis AMD Opteron and Itanium 2 servers with S2io and Intel 10-gigabyte Ethernet cards and Cisco switches, and we reached 6.8 gigabytes a second or just about 850 megabytes per second.

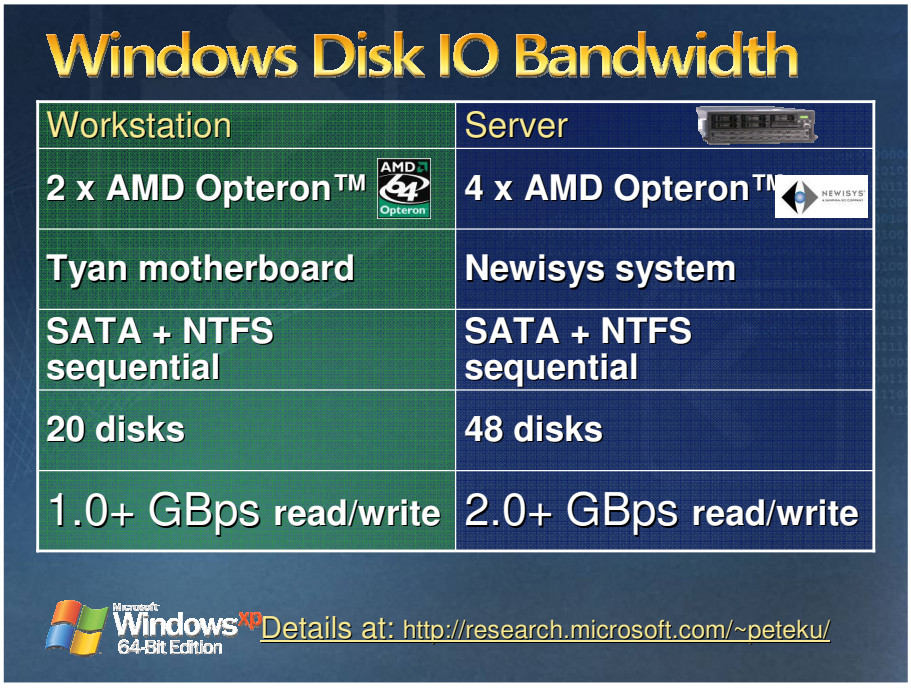
Jim:

So that's a CD per second.

Harvey:

We needed 64-bit machines for this; 32-bit Xeons top out at about 4 gigabits per second.

Now we're working towards 1 gigabyte per second.



Jim:

Harvey and I are collaborating on transcontinental disk-to-disk transfers at 1GBps.

That's the real science requirement.

Harvey moves the data 11,000 kilometers.

My job is to move it the first and last meter, from and to disk.

We measured disk bandwidth on various machines.

An NEC super-server can easily read and write at 3.5 GBps.

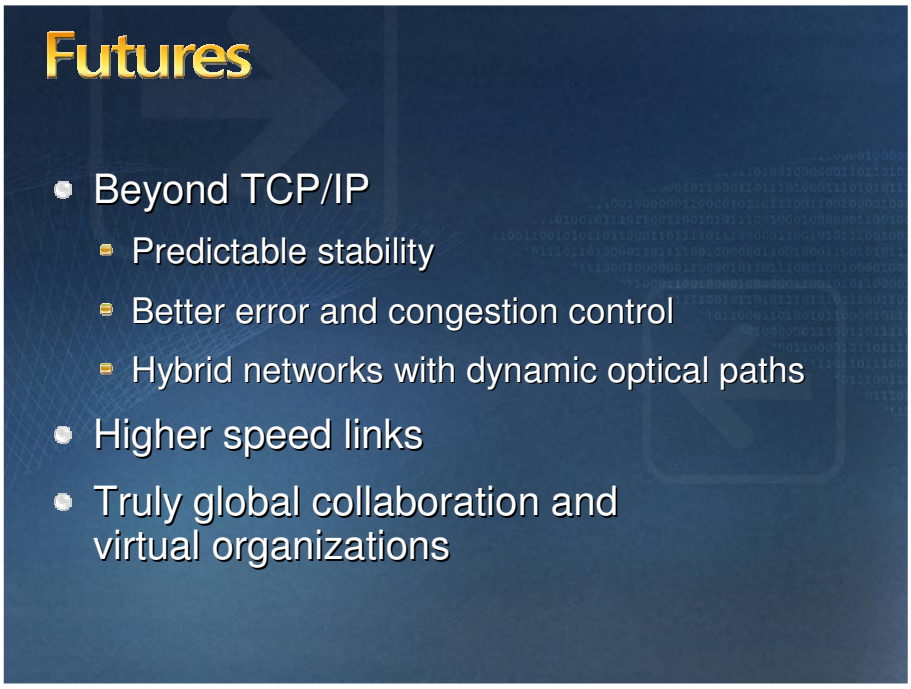
Workstation class AMD Opterons deliver about 1GBps

A Newisys 4-way AMD Opteron server delivers over 2 GBps from NTFS stripe sets.

Details of our disk measurements are on my website.

It finds NTFS on AMD Opterons deliver several gigabytes per second.

We are now working on moving data at 1GBps from a 20-disk Newisys Opteron at CERN to an identical system at Caltech.



Harvey:

I'm confident we'll soon be able to move data at one gigabyte per second or more across the globe.

But we still need a lot of work to make TCP/IP stable over long distance networks at gigabyte per second data rates.

And work is needed to put this into production use, which is what enables the science.

Data-intensive science needs these technologies to allow our global scientific communities to collaborate effectively at gigabyte per second speeds.

This is essential for the next round of discoveries at energies that were previously unobtainable.

Jim:

Networks are a lot faster than you think.

They're running at gigabyte per second speeds.

That means that you need to move things around inside the processor, IO system, and memory system, at tens of gigabytes a second in order to keep up.

Working with Caltech and working with CERN has been a wonderful thing for the Windows group.

We've learned a huge amount from them, and I really want to thank Harvey for coming here today and telling us about his work.

With that we'll turn the podium back to Bill.

Thank you very much