

Wide Area Networking@CERN

- ◆ **Status & near-term plans**
 - **Main connections (present & future),**
 - **CERN GigaPoP,**
 - **CERN USA PoP (Qwest/Chicago)**
 - **Internet routing summary**
 - **CIXP Update**
- ◆ **Evolution of Telecom infrastructure**
- ◆ **Summary**
- ◆ **Technical challenges**
 - **QoS**
 - **very high speed file transfer**
- ◆ **GEANT**
- ◆ **WEB100**

Main Internet connections@CERN (1)

- **RENATER (French Academic & Research Network).**
 - **Mostly for CEA (Saclay) private use**
- **SWITCH Next Generation (Swiss Academic & Research Network):**
 - **2.5 Gbps pilot between CERN & ETH Zurich early 2001**
 - **DWDM over dark fibers rented from Intelcom (end 2001)**
 - **will make it easy to build lambda based VPN**
- **TEN-155 (Trans-European Network, 155 Mb/s).**
 - **Combined CERN-SWITCH access (25% CERN, i.e. 40Mbps)**

Main Internet connections@CERN (2)

- **US Line consortium (USLIC)**

- CERN, US/HEP (via Caltech & DoE), Canada/HEP (via Carleton)
- IN2P3 (CCPN Lyon).
- World Health Organization (WHO).

- **JAPAN**

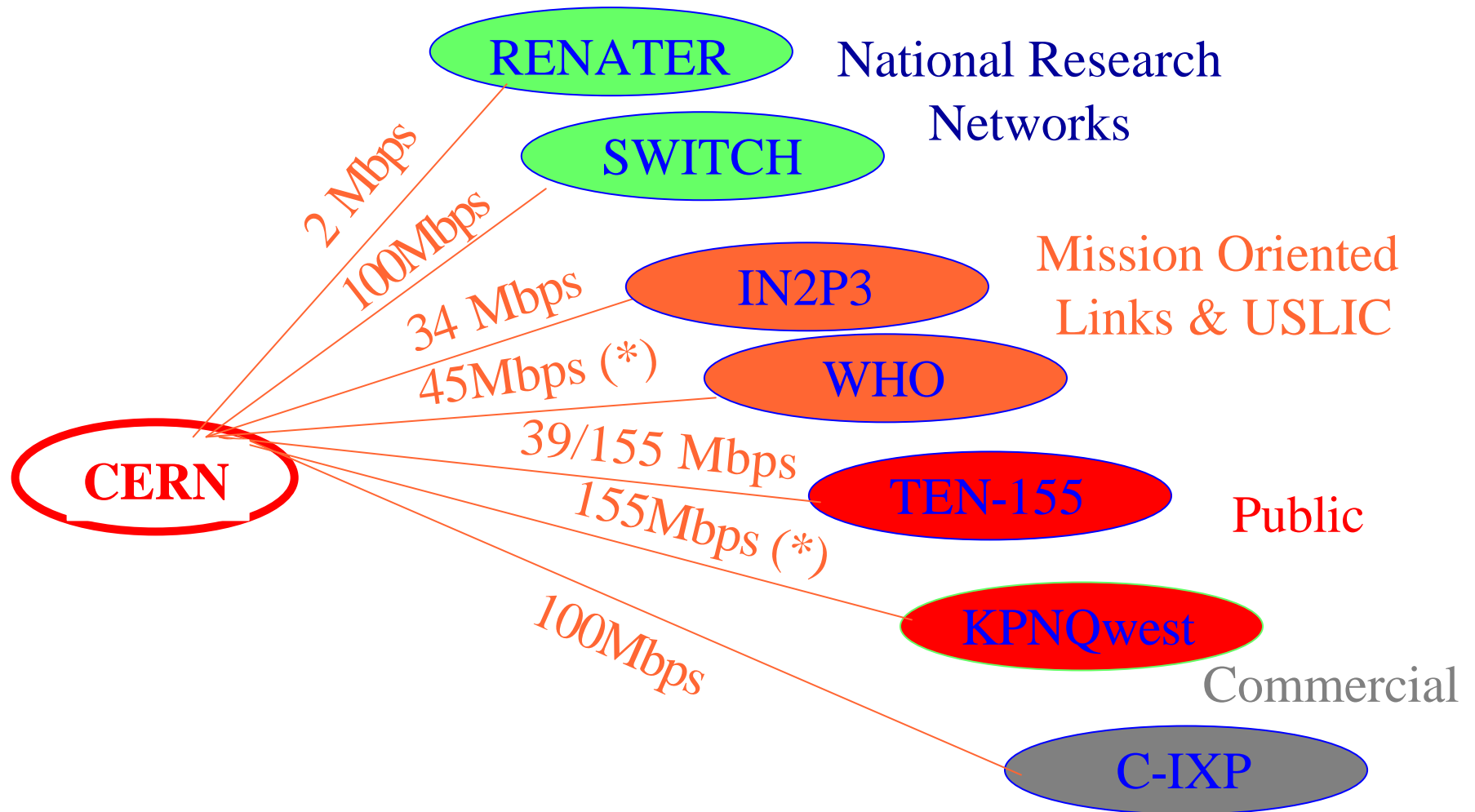
- **NACISIS (4Mbps ATM/VP over TEN-155's Managed Bandwidth Service (MBS))**

- **CIXP (CERN Internet Exchange Point)**

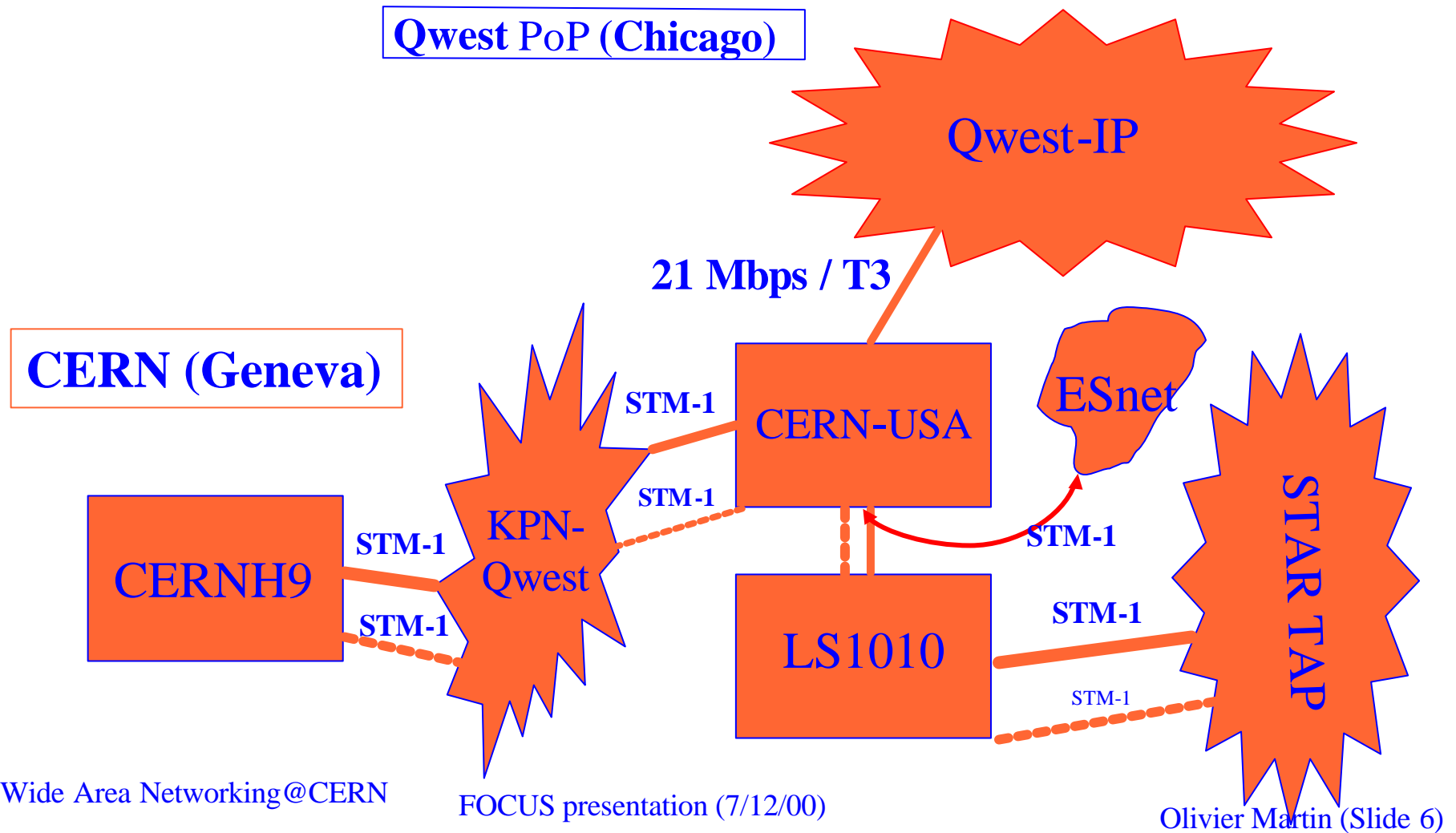
Near term plans

- **GEANT (Gigabit European Academic NeTwork, 2.5 Gbps core initially).**
 - expect arrangement, similar to the one in place for TEN-155, to continue with SWITCH:
 - shared 1Gbps access (timeframe: Summer 2001)
 - CERN purchasing up to 2*155Mbps (i.e. 25%) and Switch the rest.
 - Then double access bandwidth every 12-18 months.
 - CERN may host the Swiss PoP of GEANT, the successor of TEN-155 to be deployed around the middle of year 2001:
 - would make it easier to access new services (e.g. lambda)
- **USLIC**
 - 45M circuit to Chicago and STAR TAP to be upgraded to 155M (STM-1)
 - timeframe: mid-December 2000
 - may exercise the option to configure as 2 unprotected SDH circuits in order to:
 - allow more innovative options to be used while maintaining the overall stability of the service

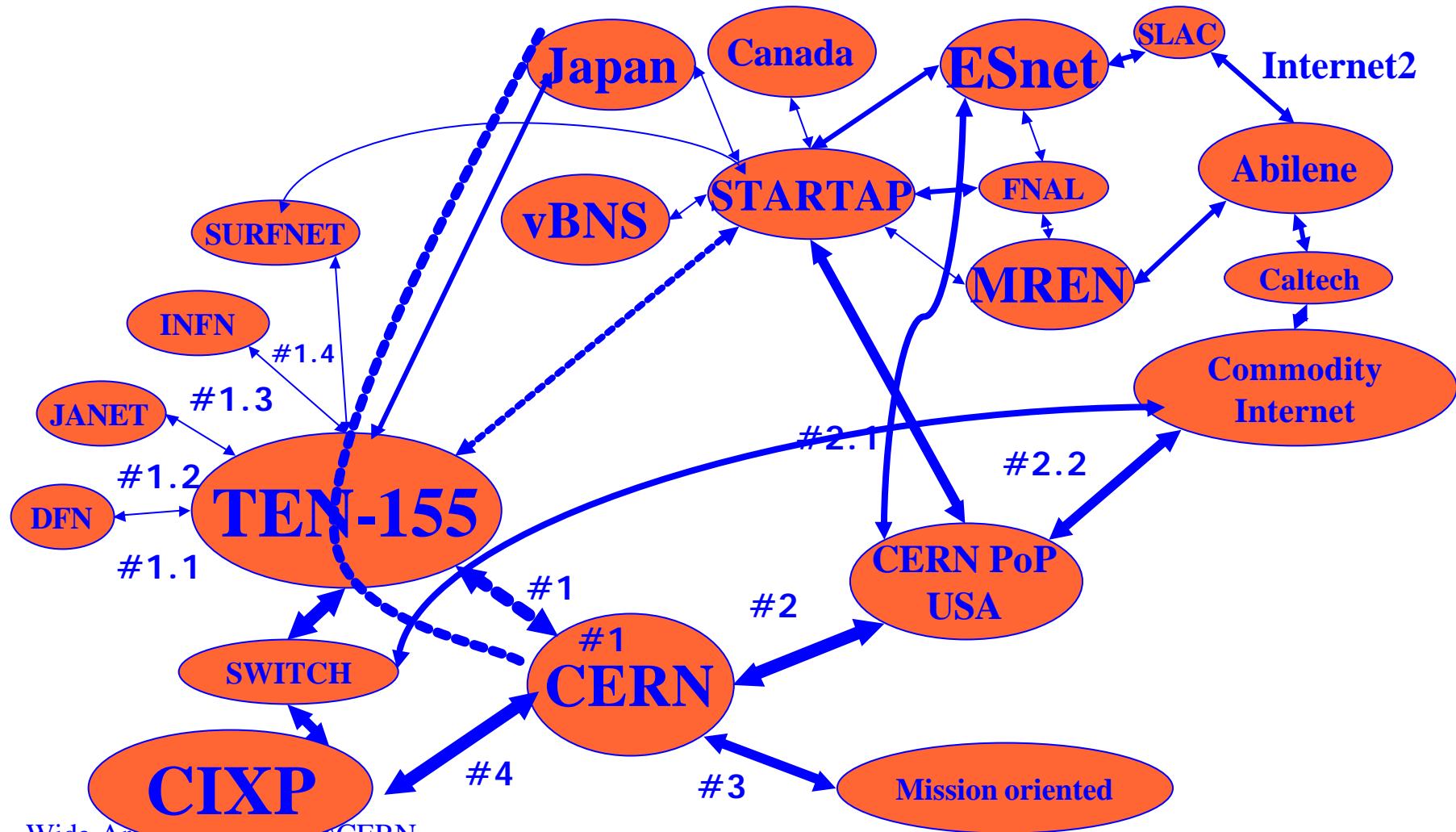
CERN GigaPoP (December 2000)



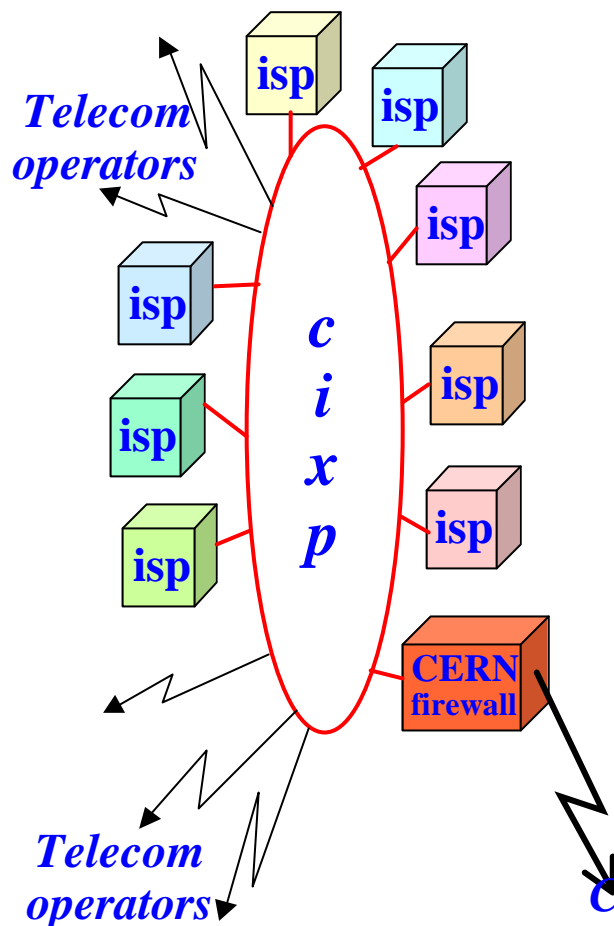
KPNQwest co-location plan (Chicago)



Internet Access@CERN



CERN Internet Exchange Point (CIXP) (December 2000)



Wide Area Networking@CERN

Telecom providers:

Cablecom, COLT, diAx, France Telecom, GTS, KPNQwest (*), LTT(*), Multilink, MCI/Worldcom, SIG, Sunrise, Swisscom (Switzerland), Swisscom (France), Thermelec.

Internet Service Providers:

Infonet, AT&T Global Network Services (formerly IBM), Cablecom, C&W, Carrier1, Colt, Deckpoint, Deutsche Telekom, diAx (dplanet), EBONE, Eunet/KPNQwest, France Telecom OpenTransit, Global-One, Globix, HP, **InterNeXt**, ISDnet/Ipergy, IS Internet Services (ISION), LTT(*), Net Work Communications (NWC), PSI Networks (IProlink), MCI/Worldcom, Petrel, Renater, Sunrise, Swisscom IP-Plus, SWITCH, TEN-155, Urbanet, VTX, Internet Network Services (Wisper/INSnet).

Telecom infrastructure@CERN (1)

◆ 1984-1989

- EARN/BITNET era: 9.6 --> 64 Kbps
- Public X.25 via Swisscom at 48Kbps

◆ 1989-1997

- Start of the LEP era, first 2Mbps A&R circuit between CERN and Italy (INFN/CNAF (Bologna)).
- 4 pair of fibers (Swisscom)
 - » saturated end 97 (2*140Mbps, 2*155 Mbps ATM)
- second optical fiber cable & redundant SDH 622Mbps loop.
- France Telecom (still on old copper fibers)

Telecom infrastructure@CERN (2)

- ◆ **1998 onwards (Start of Telecom de-regulation era):**
 - **SIG (Services Industriels de Geneve)**
 - » Thermelec, diAx, Sunrise, MCI / Worldcom, GTS, Cablecom, Multilink (DT)
 - **France Telecom**
 - » 2 fiber optic cables, operational 2.5 Gbps SDH loop, DWDM equipment to be installed 1Q01.
 - **COLT, KPNQwest (*) & Swisscom - (France) also come with their own cables.**
- ◆ **Number of Optical Fiber cables:**
 - **Swisscom/CH (2)**
 - **France Telecom (2)**
 - **Colt (2)**
 - **KPNQwest (2)**
 - **SIG (4 cables with 144 fibers each)**

 - **Swisscom/FR(1).**
- ◆ ***Estimated number of strands of fibers: 800 (i.e. potentially unlimited bandwidth at hand)***

Telecom infrastructure@CERN (3)

- ◆ **Several SDH 2.5 Gbps loops in operation (10Gbps links are just starting to appear in Europe & USA).**
 - **Provisioning new circuits becomes a very easy and also very quick process.**
- ◆ **Geneva, and in some cases CERN, is now part of most new emerging pan-European backbones (e.g. KPNQwest (*), Swisscom, Colt, France Telecom, GTS).**
- ◆ **Would not have happened without the presence of the CERN Internet eXchange Point (CIXP):**
 - **The CIXP is in the process of being distributed over dark fibers and Gigabit Ethernet links to several neutral expansions sites in the Geneva area.**
 - » ***The extension to Telehouse Geneva 2*Gbps Ethernet is operational.***
 - **The CIXP is not limited to peerings between ISPs, it is also used by Telecom Operators to interconnect**

Future prospects

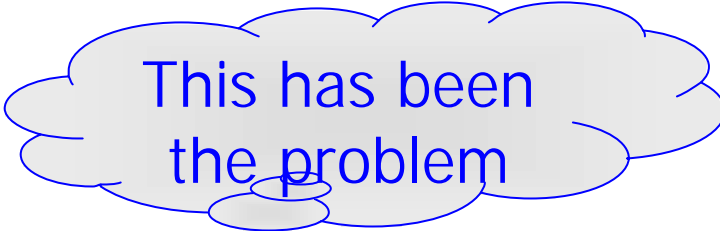
- ➔ WDM technology evolving very fast,
 - ➔ Coarse WDM, Dense WDM, Hyperfine WDM (~4000 channels)
- ➔ New pan-European backbones, as well as new transatlantic cables, are being deployed
- ➔ Prices are falling down very rapidly, therefore, we can be reasonably confident that:
 - ➔ **we will, at least, double our external networking capacity each year.**
- ➔ The combination of prices going down and (maybe) a moderate budget increase (20-25%) may make it possible to move towards multiple STM-4 (622Mbps) or even STM-16 (2.5Gbps) faster than originally planned (i.e. before 2005):
 - ➔ **But this only makes sense if there are real prospects to make effective use of the capacity “end to end”, which is far from being the case today.**

WAN vs LAN bandwidth

- ◆ **The common belief that WAN will always be well behind LANs (i.e. 1-10%) is, according to me, wrong but admittedly controversial.**
 - *WAN technology is well ahead of LAN technology, state of the art is 10Gbps (WAN) against 1Gbps (LAN)*
 - *Price is less of an issue as they are falling down very rapidly.*
 - *Some people are even advocating that one should now start thinking new applications as if bandwidth was (almost) free!*
 - *This may sound a bit premature, at least, in Europe, however, there are large amounts of unused capacity floating around!*

Quote from Jim Gray Microsoft Research

- ◆ **WANS are getting faster than LANS**
G8 = OC192 = 10Gbps is “standard”
- ◆ **Link bandwidth improves 4x per 3 years**
- ◆ **Speed of light (60 ms round trip in US)**
- ◆ **Software stacks**
have always been the problem.
- ***Time = SenderCPU + ReceiverCPU + bytes/bandwidth***



This has been
the problem

Is there a WAN bandwidth glut in sight?

◆ European situation:

- bandwidth glut is real,
 - » however, local loops are still a problem
- lambda service are becoming available,
- dark (actually dim) fibers are still a relatively scarce resource outside MAN environments.
 - » SWITCH is the only known example of NRN migrating from Telecom operated to community managed dark fiber backbone network.

◆ Transatlantic situation:

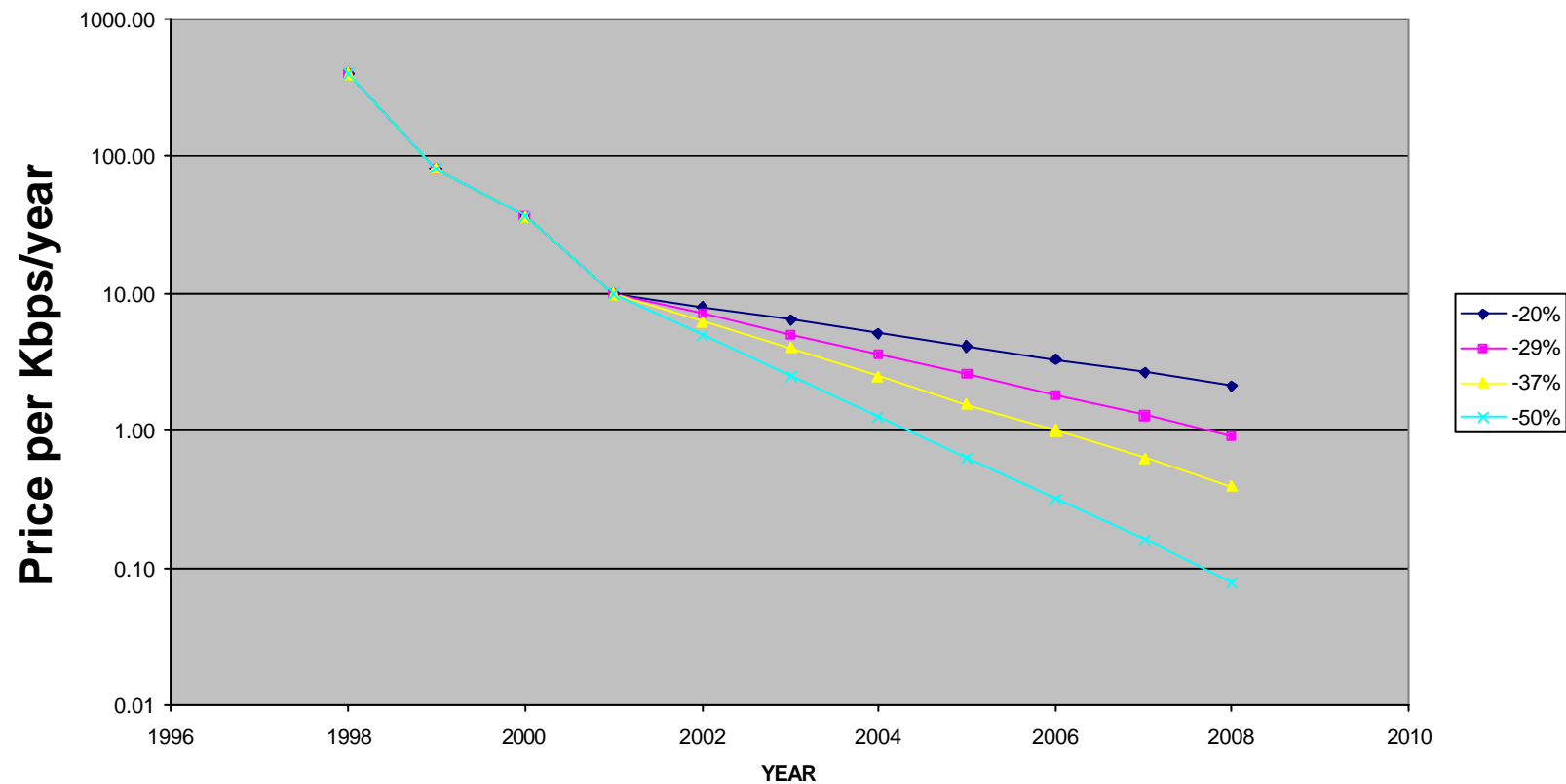
- the relative bandwidth glut is also real, but
 - » how long will it last?
- lambda service are becoming available,
- dark fibers are almost unthinkable given the rather small number of fibers per cable.

So, where does that lead us?

- ◆ **Future clearly lies with "all" optical networks.**
 - entering new era
- ◆ **Conventional Telecom Provider model, i.e. managed bandwidth, SDH, etc, increasingly challenged, especially in Canada.**
- ◆ **Strong advocates (Canarie project) of dark fiber based networks:**
 - with CWDM & 10GBE technology,
 - and, still to be developed, Optical Cross-connect coupled with Optical BGP capable routers in order to:
 - » control allocation of wavelength
 - » provide end to end QoS.

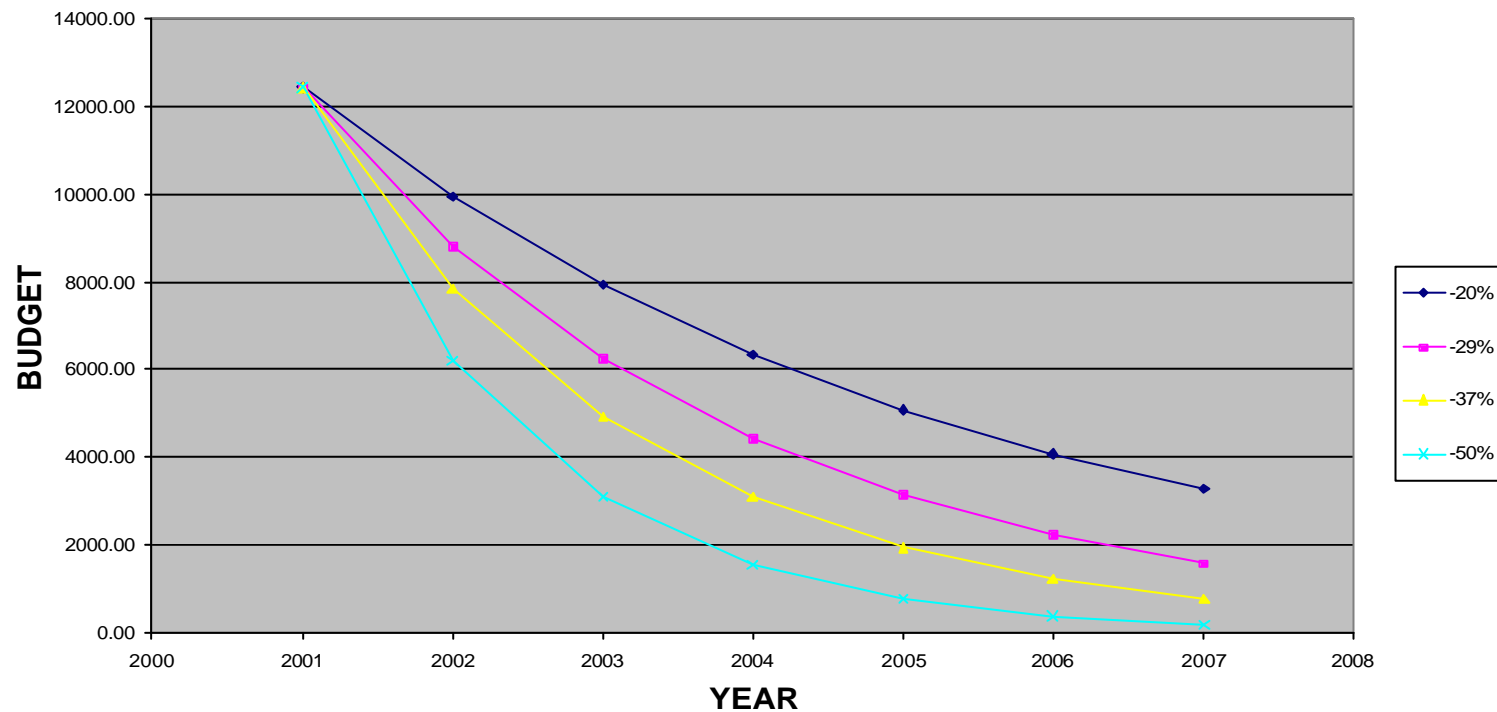
Evolution of Transatlantic circuit prices

Evolution of Transatlantic prices



2.5 Gbps Transatlantic circuit price estimates

2.5 Gbps costs (hypothesis 8*STM-1)



SUMMARY

- ◆ **Multiple circuits from CERN to Tier1 regional centers at up to 2.5 Gbps (i.e. STM-16/OC-192c) will be possible by 2003-2005.**
- ◆ **Cost may be problematic (1-3MCHF per circuit).**
- ◆ **Very high speed LANs implied.**
- ◆ **QoS & Gigabit/second file transfer on large bandwidth*delay paths may still be problematic.**
- ◆ **The public Internet as well as national research networks are evolving in a way nobody can predict.**
 - **This will have a profound impact on the LHC**

Technical Challenges

- ◆ **QoS deployment**
- ◆ **Very high speed file transfer**
 - **pre-requisites**
 - **problems**
 - **tools**
 - **tuning**
 - **single vs multiple streams**
 - **results**

QoS Deployment

- ◆ **CAR capable IOS versions installed, some bugs identified, capability disabled.**
 - *Still waiting for “the” stable release with the latest features (e.g. DSCP marking), but sounds like a moving target!*
 - *negative impact of diffserv policing on TCP/IP congestion handling mechanisms should not be underestimated.*
- ◆ **QoS mechanisms needed for:**
 - **VRVS (Virtual Room Videoconferencing System)**
 - » *Planring to use RSVP and/or diffserv*
 - **IP telephony (CERN, DESY, FNAL, SLAC)**
 - » *Priority queuing adequate*
 - **Video on Demand services**
 - » *IGRID'2000 demos with iCAIR (diffserv capable IBM Video Charger)*
- ◆ **QoS authenticated Umich proposal (Merit, Internet2/UCAID, ATLAS, CERN)**
 - **vic & vat extensions to interact with interdomain authenticated bandwidth brokers and diffserv expedited forwarding.**

QoS Monitoring

◆ Tools:

- Ping, Traceping
- RIPE, NIMI & Surveyor probes installed.
- Netperf, Iperf & Tcptrace used for benchmarking & tuning
 - » monitoring throughput has proven to be more fundamental than expected!
- Monitoring Web performance
 - » url-get
- Real-time flow visualization & consolidation using custom program (Jiri Navratil) instead of cflow.
- statistics collected with SNMP polling and Netflow.

◆ CERN's external networking statistics are available from:

<http://sunstats.cern.ch/mrtg>

Very high speed file transfer (1)

- ◆ **Let us start with the obvious prerequisites first:**
 - **High performance switched LAN required:**
 - » *requires time & money.*
 - **High performance WAN also required:**
 - » *also requires money but is becoming possible.*
 - » *very careful engineering mandatory.*
 - **Make sure the (strong) security architecture (e.g. firewall, encryption) does not conflict with the high throughput requirements.**
 - **Monitor achievable file transfer throughput in order to get some feeling for what can be achieved in practice.**
 - » *under certain hypothesis, can also be derived from Ping packet loss rates.*

Very high speed file transfer (2)

- ◆ **Keep aware that tcp/ip does not perform well over large Bandwidth*Delay Paths (BDP):**
 - **Performance inherently unstable!**
 - **Bbftp, i.e. Multi-stream (a la Babar with compression) vs single stream, does it help?**
- ◆ **Tcp relays (Performance Enhancing Proxies)**
 - **General idea, as things get worse with distance, try to split the problem:**
 - » at source/destination
 - » in the middle
 - » on the fly (i.e. transparent) vs store&forward (a la Interplanetary Internet gateways)
 - » advantages, easier tuning in a well controlled environment (e.g. choice of best suited hardware/software platform)
 - » disadvantages: single point of failure(s), scaling
 - » Does it make sense?

◆ **Can cache and/or CDN technology help?**

Very high speed file transfer (3)

◆ Tools

- **tcptrace/xplot**
- **tcpillust**

◆ Tuning:

- **Very delicate, lot of parameters:**
 - » window scale
 - » SACKs/FACKs
 - » Timestamps
 - » delayed ACKs

◆ Some understanding of tcp/ip operations regime required:

- » slow start ($cwnd < ssthresh$),
 - ◆ **$cwnd += SMSS * SMSS / cwnd$**
- » congestion avoidance ($cwnd > ssthresh$),
- » fast retransmit & fast recovery

Very high speed file transfer (4)

*TCP's "congestion avoidance" algorithms are not compatible with high speed, long distance networks. The "cut transmit rate in half on packet loss and increase the rate additively" algorithm will simply not work. Consider a 10Gbps link to a destination half way around the world. A packet drop due to link errors (not congestion or infrastructure products) can be expected about every 20 seconds. However, with a RTT of 100ms (not even across the continent), if a TCP connection is operating at 10Gbps, the packet drop (due to link error) will drop the rate to 5Gbps. It will take 4 *MINUTES* for TCP to ramp back up to 10Gbps.*

- ◆ *Therefore, there needs to be a change to TCP's congestion avoidance algorithm for future high speed networks. Matt Wakeley (Agilent)*

If you want to stay in the regime of the TCP equation, 10 Gbit/s for a single flow of 1460 byte segments at 100 ms (quarter-earth) RTT means you need a packet loss rate of about $1E-10$, i.e. you should lose packets about once every five hours.

- ◆ *Clearly, this kind of congestion signal is way too coarse for any useful control algorithm. Dr. Carsten Bormann, University of Bremen*

Very high speed file transfer (5)

- ◆ **tcp/ip fairness is a myth,**
 - **does not quite work over large bandwidth*delay path (BDP) because of the 1 MSS per RTT congestion avoidance algorithm.**
- ◆ **Web100, a 3MUSD NSF project, might help enormously!**
 - » better TCP/IP instrumentation (MIB)
 - » **autotuning**
 - » tools for measuring performance
 - » **improved FTP implementation**

◆ **The GEANT project**

- **Main elements of the ongoing Call for Tender.**
- **Services**
- **Status**
- **CERN connectivity**

GEANT Project Update

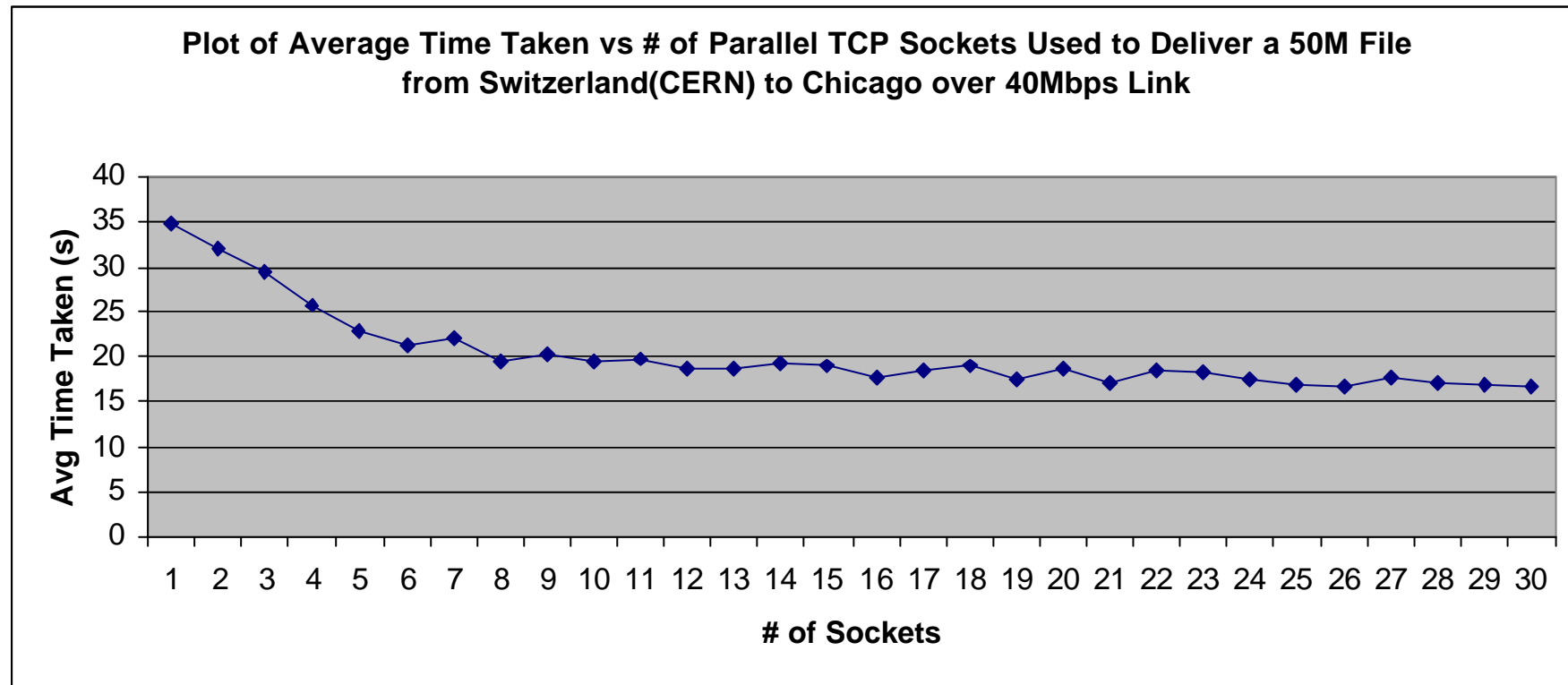
◆ STATUS

- ◆ Call for Tender underway (closing date: September 29)
- ◆ 3 months negotiation period
- ◆ 3-12 months implementation period
- ◆ core backbone and first links expected during 3Q01.
- ◆ Transition from TEN-155 to GEANT to be completed by the end of 2001.
- ◆ Surfnet situation still unclear but there are encouraging signs that they will join eventually.

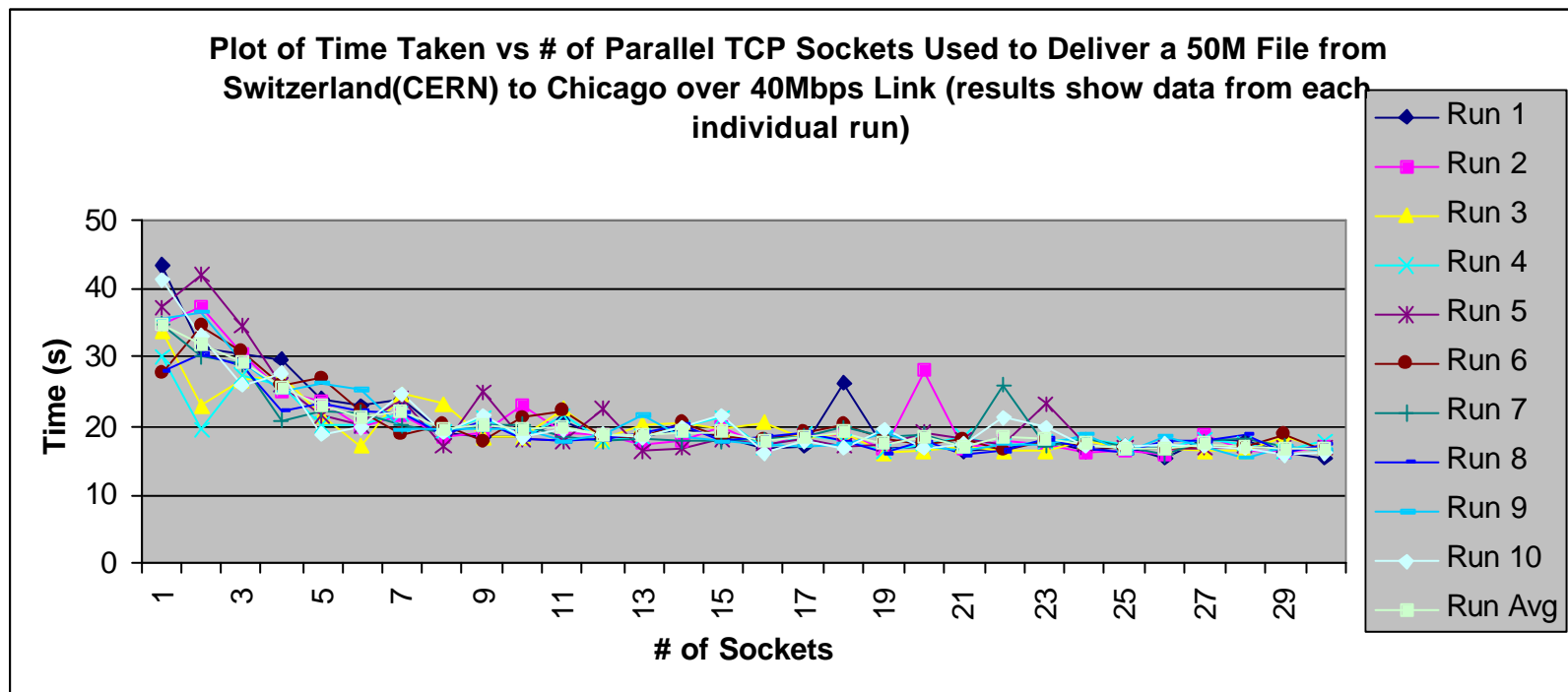
◆ CERN

- ◆ shared 1Gbps access with SWITCH with up to 2*155Mbps dedicated to CERN depending on exact pricing (fixed budget).

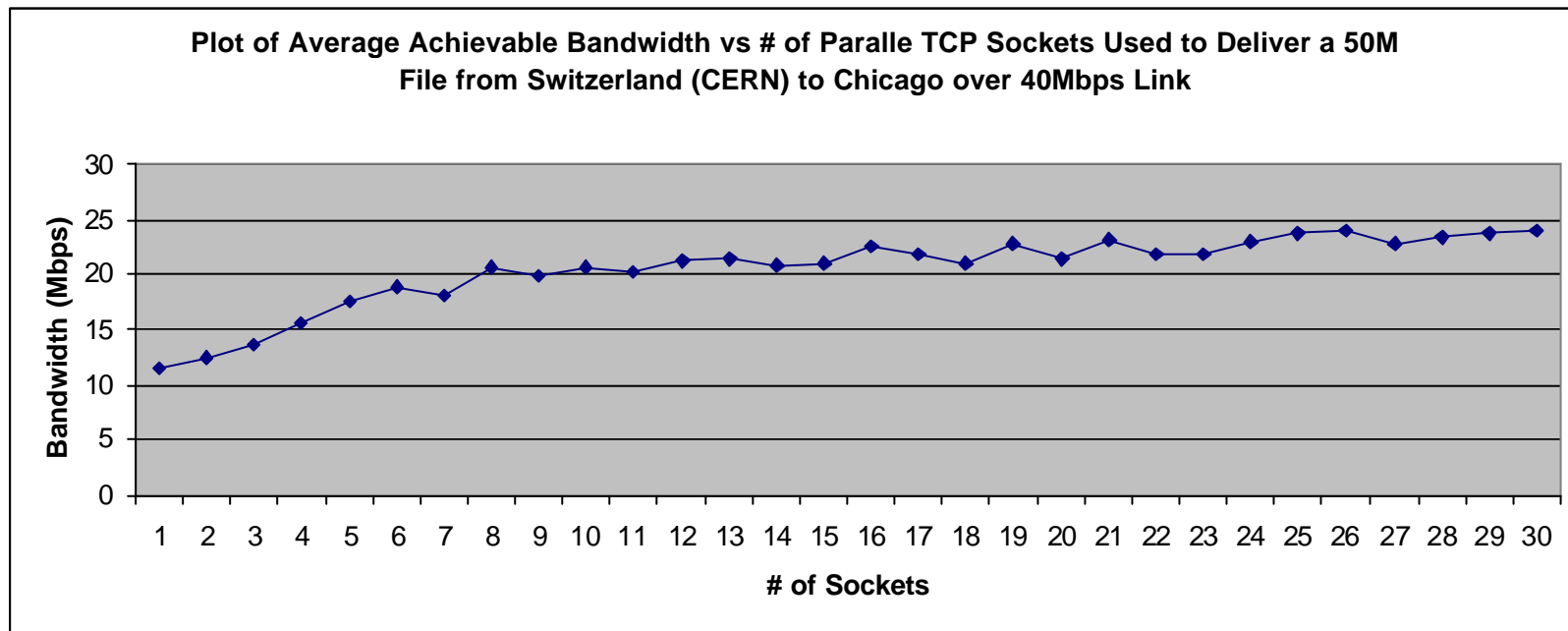
Throughput Monitoring (CERN--->UIC/EVL)



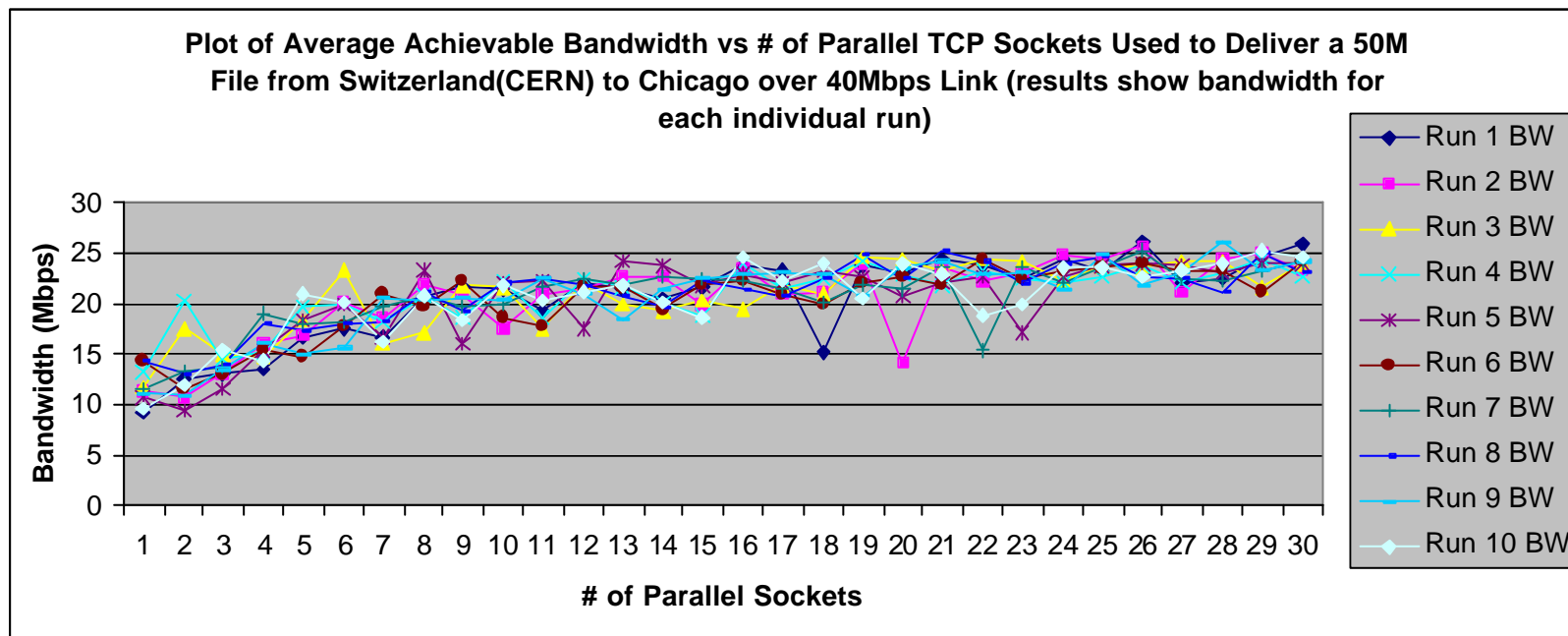
Throughput Monitoring (CERN--->UIC/EVL)



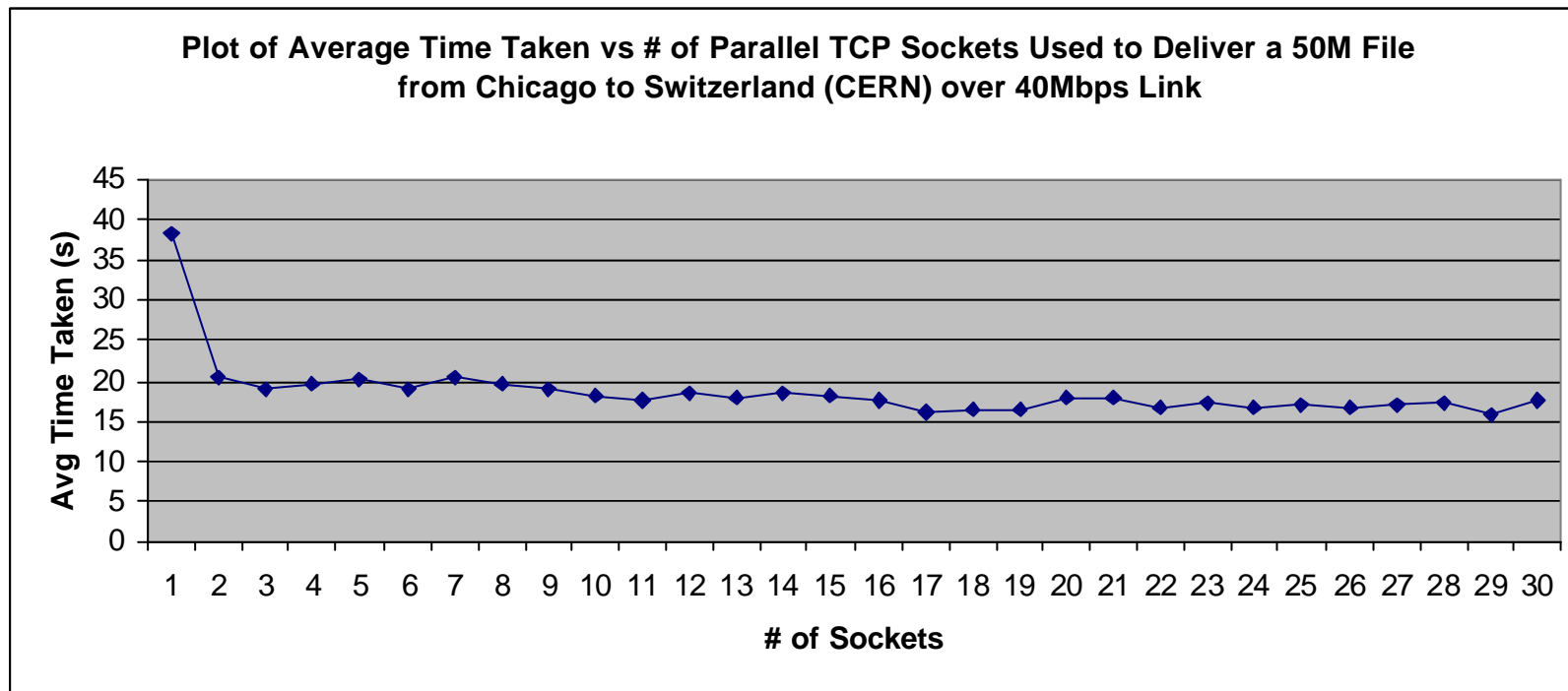
Throughput Monitoring (CERN--->UIC/EVL)



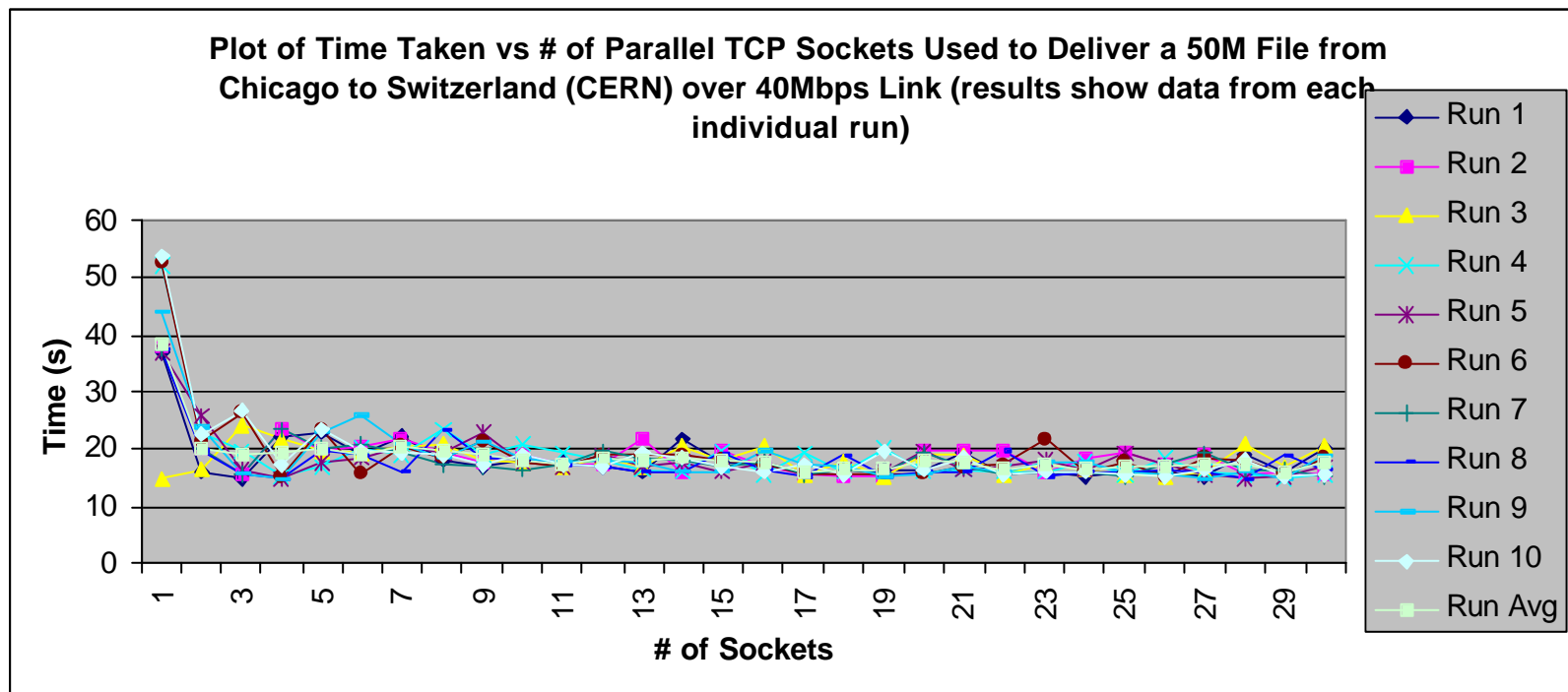
Throughput Monitoring (CERN--->UIC/EVL)



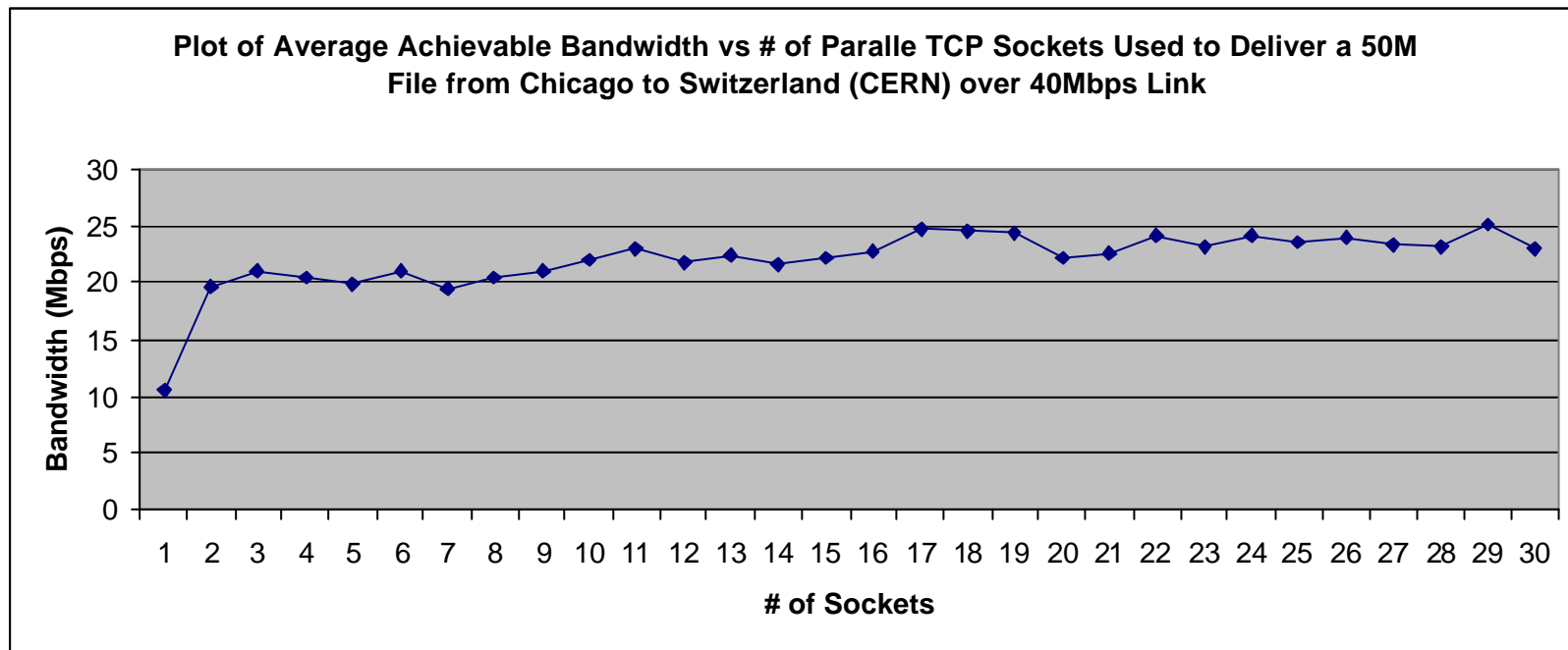
Throughput Monitoring (UIC/EVL--->CERN)



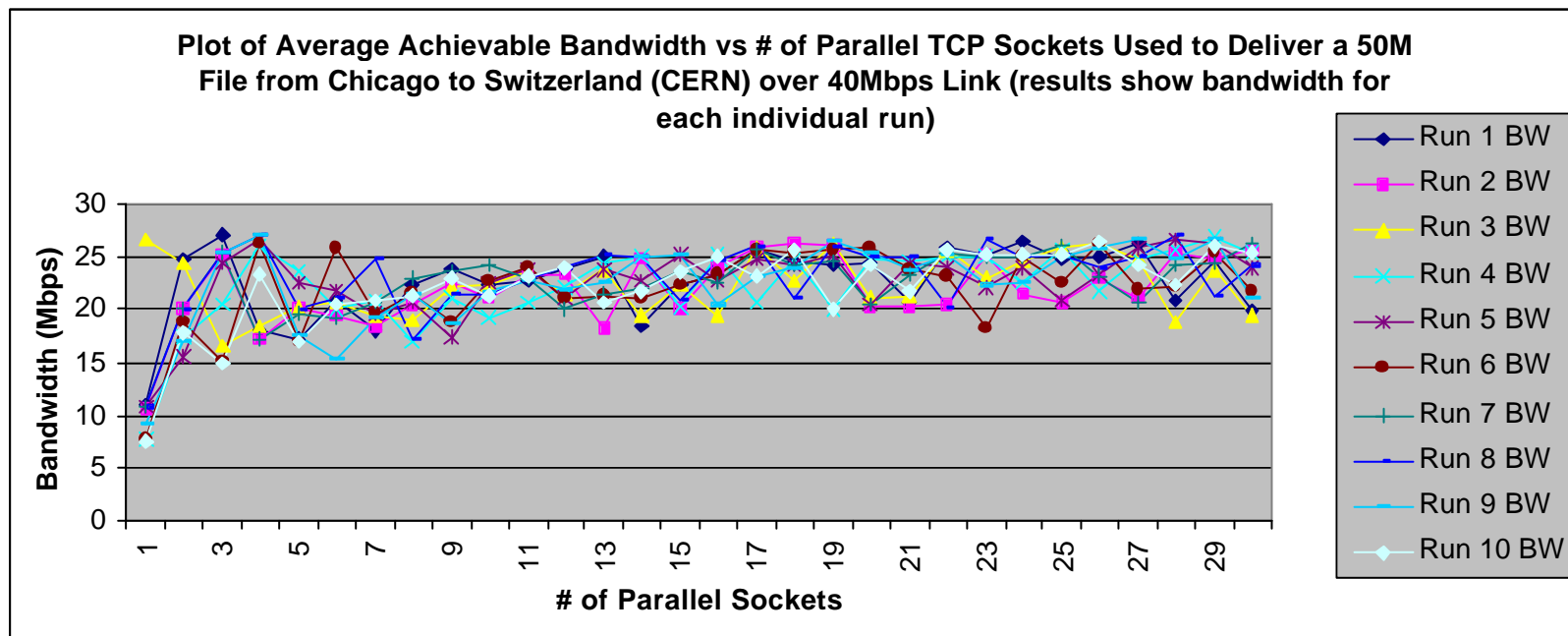
Throughput Monitoring (UIC/EVL--->CERN)



Throughput Monitoring (UIC/EVL--->CERN)



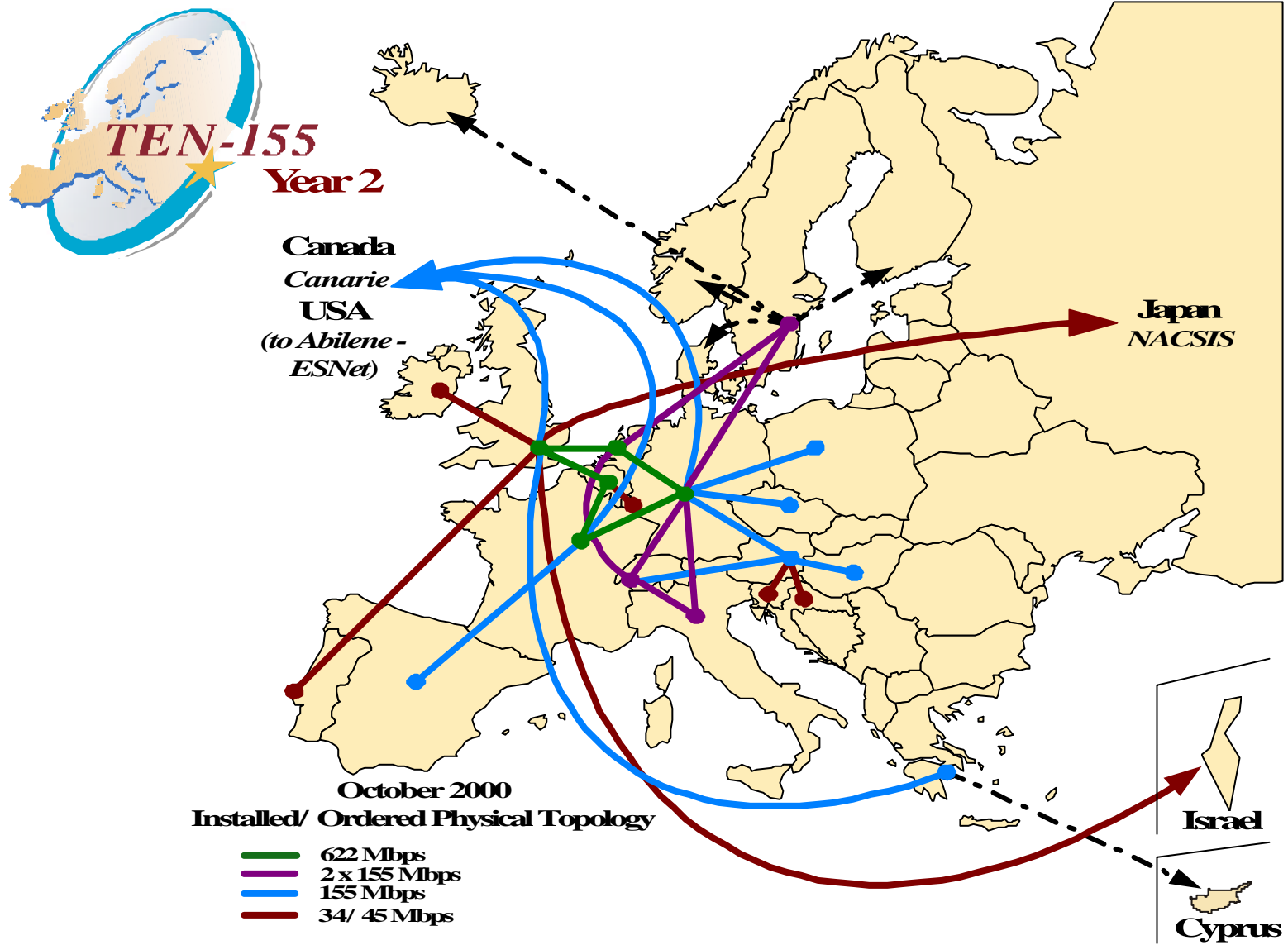
Throughput Monitoring (UIC/EVL--->CERN)



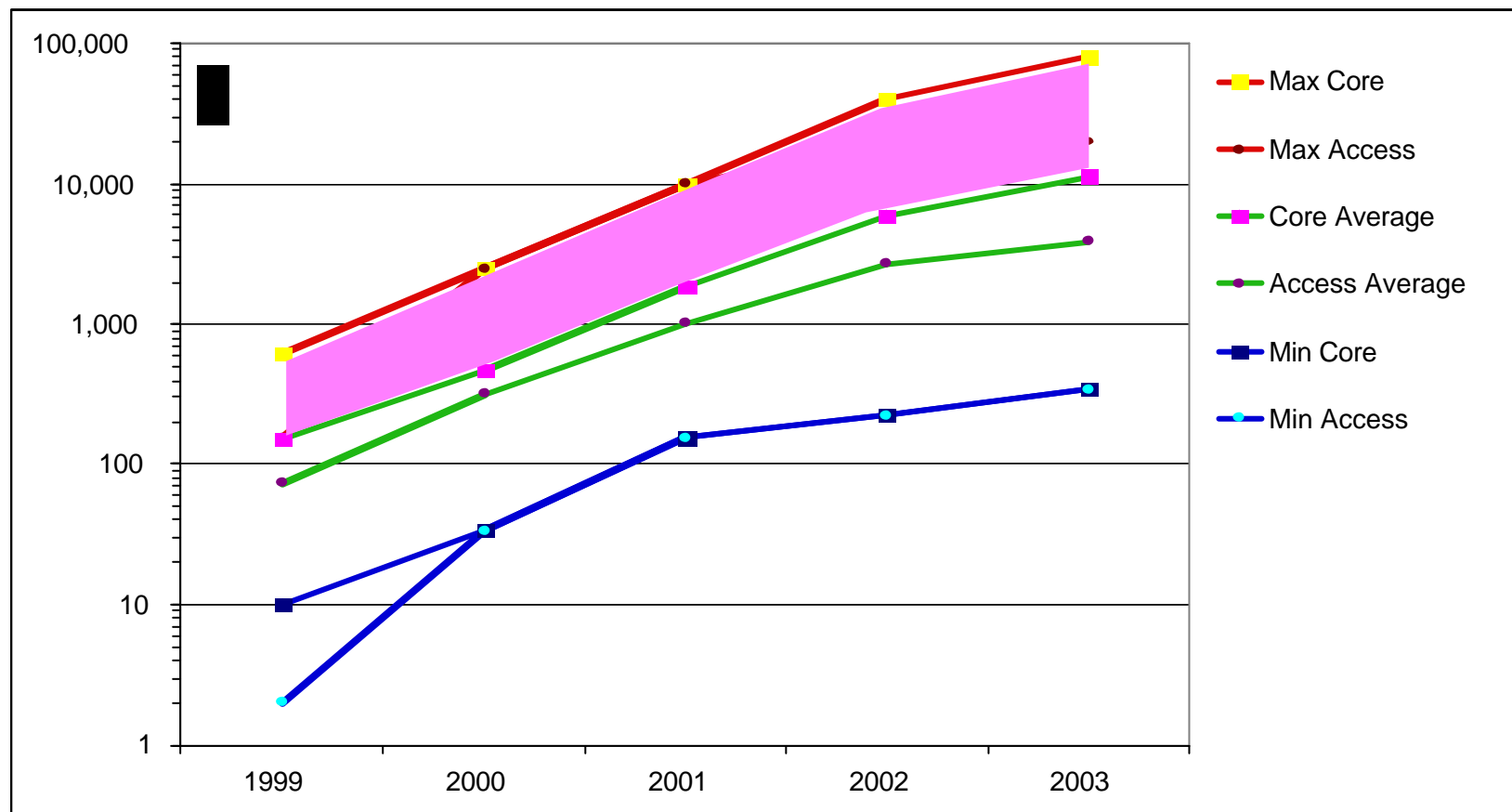
European Gigabit Network

Fernando Liello

European NREN Consortium



Bandwidth Requirements



GEANT Services

- ◆ **Basic IP service**
 - **including connection to research networks in other regions**
- ◆ **Native multicast**
- ◆ **Premium IP service**
- ◆ **Guaranteed capacity**
- ◆ **Virtual private networks**
- ◆ **New services**
 - **IPv6**
 - **support for “disruptive” testing (not ATM)**

European Distributed Access (EDA)

- ◆ **Rationalize Research Connections With Other World Regions**
 - **N.–America, Asia–Pacific, S.–America, ...**
- ◆ **Connections May Be Activated at Any PoP of the Core**
- ◆ **Transparent Access To/From Any Other PoP**

“Vision”

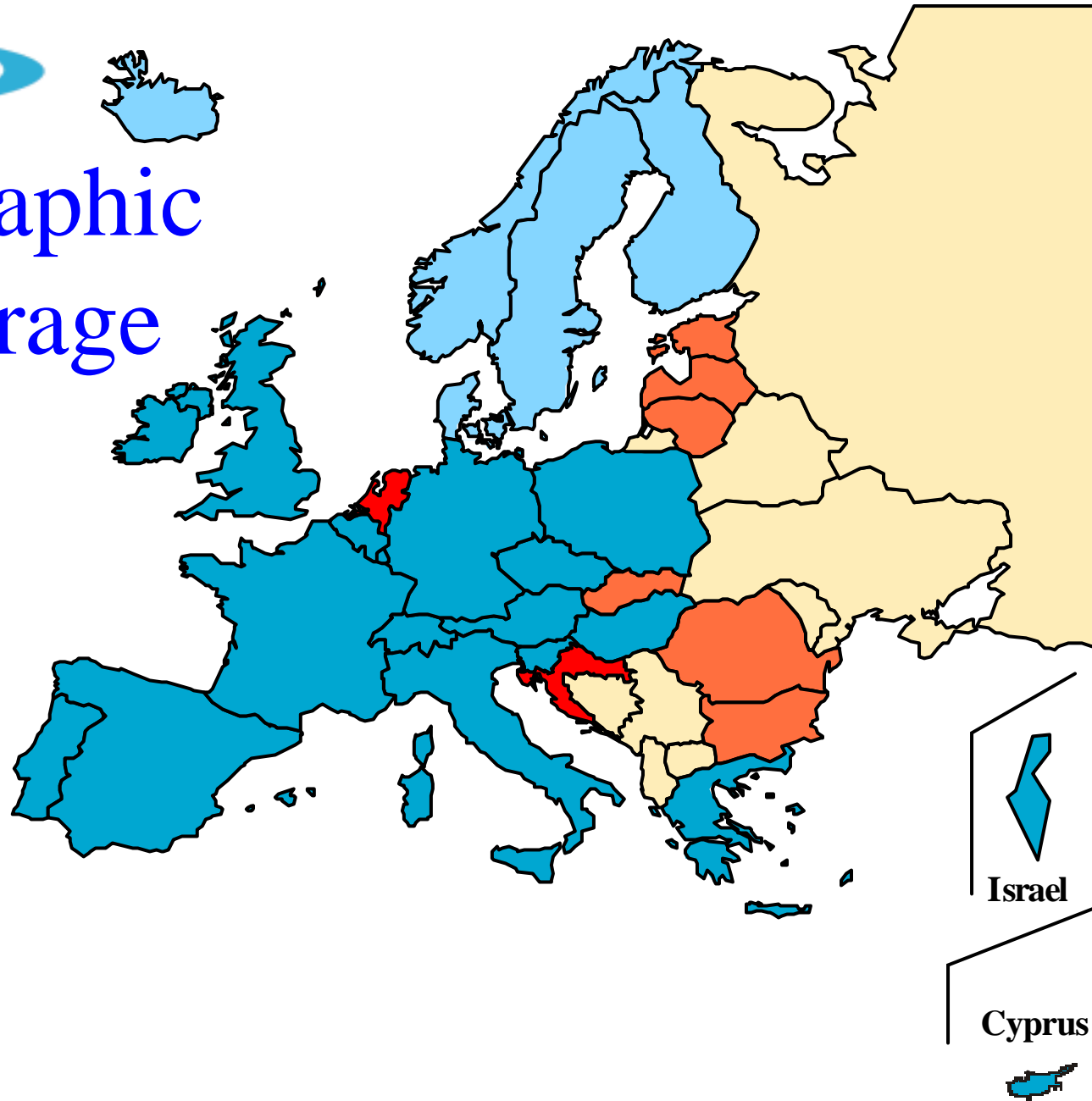
- ◆ **ideal is dedicated fibre + DWDM, but**
- ◆ **is it available (throughout Europe)?**
- ◆ **is it ready (for operational service)?**
- ◆ **is it affordable?**
 - **equipment costs?**
 - **management in so many countries?**

GEANT Reply

- ◆ **4 years plan**
- ◆ **About 220 M€budget**
 - **International (within Europe) & EDA support**
 - **Contract among EC, 25 NREN's & DANTE**
- ◆ **Intercontinental (Asia–Pacific, S. & N. America, Mediterranean) Connectivity Budget TBD during 2001**
 - **Co–funding agreements being negotiated**
 - **EC calls during 2001**
- ◆ **Official Launch: 6/11/00 at IST–2000 in Nice**



Geographic Coverage



Will Capacity Goals Be Met?

- ◆ **Initial Speeds 2.5 - 10 Gbps?** **yes**
- ◆ **Direct Access to Wavelength Capacity 6-10 Locations Initially?** **yes**
- ◆ **Direct Access to Wavelength Capacity 20+ Locations within 4 years?**
 Very probable
- ◆ **100 Gbps capacity within 4 years?**
 Very probable

Tests/Piloting

◆ Year 1 commitment

- premium IP
- MPLS/VPN
- multicast improvement
- IPv6
- traffic measurement

◆ Annual update of “technology roadmap”

Summary

- ◆ **The Most Advanced Possible Networking Support for the Research and Education Community Throughout Europe**
- ◆ **A High Quality Infrastructure for World–wide Research Connectivity.**

Web100

Basil Irwin & George Brett

Web100 Project

What Is Web100?

- ◆ **A Software Suite which would:**
 - **Enable ordinary network users to attain full network data rates across all networks without requiring help from networking experts.**
 - **Automatically tune the network stack for underlying network environment.**
 - **Provide network diagnostic and measurement information.**

Motivation

From the July 1999 report from the *Advanced Networking Infrastructure Needs in the Atmospheric and Related Sciences (ANINARS) Workshop*

“ FTP (or FTP-like) bulk data-transfer is the most important networking function used to construct applications in this scientific discipline, yet failure to achieve effective bandwidths equal to apparently available bandwidths is most evident with bulk data-transfer applications. A variety of host-software problems contributes to this failure, and [NSF] programs should be developed to help solve these problems.”

Motivation

- ◆ **Applications, researchers, end-hosts are still unable to fully utilize high performance network infrastructure.**
 - **Still common for end-user to be unable to exceed 3 Mbps without help from a networking expert.**

Why Poor Performance ?

- **Network software (mostly TCP) optimized for low bandwidth -- not high bandwidth --environments.**
- **Lack of effective instrumentation and tools to diagnose end-host performance issues.**

Web100 is designed to address both these issues

Who

- ◆ **Any Application that transfers large data sets.**
 - **Application Independent**
 - **Example Areas**
 - » Atmospheric Sciences
 - » Astronomy
 - » High Energy Physics
 - » Biomedical
 - » Geo Sciences

Technical Description

- ◆ **Three critical areas:**
 - **Core Technologies**
 - **User Support**
 - **Vendor Liaison**

In order for the core technologies to be successful, they must be usable by the end-user and adopted by the vendor community.

Core Technologies

◆ OS Instrumentation

- **Instrumentation of TCP using a MIB (Management Information Base) is the foundation of Web100**
 - » Real-time per-TCP-session metrics
 - » Allow Autotuning within a host
 - » Identify bottlenecks between sender and receiver.
 - » Support measurement for end-user performance.
 - » Diagnostics for network engineers and system administrators.

Initial (Network) Applications

- **Autotuning**
 - » automatically adjust TCP to achieve maximum throughput across all connections within the host.
 - » Will use TCP-MIB to get information and then set parameters for automatic, transparent and dynamic tuning of TCP stack.
- **Performance Measurement Tools**
 - » Suite of host based measurement tools using TCP-MIB.
 - ◆ Real-time graphical display of application performance, diagnosis of network path, etc.

Interim Solution

- **FTP**
 - » Optimized FTP by leveraging ongoing application tuning work at NCSA.
 - » Will not initially use Web100 TCP-MIB
 - ◆ uses existing socket API to adjust TCP tuning parameters to support the estimated delay bandwidth product.
 - » Provides necessary stop gap measure
 - ◆ Short-term: while TCP-MIB is being developed.
 - ◆ Long-term: for non-Linux users until Web100 methods are adopted by proprietary vendors

Progress To Date

- **Prototype TCP-MIB**
 - » Has been designed.
 - ◆ 80+ Variables have been defined
 - Includes some (~6) Read-write Variables.
- **Web100 Software**
 - » TCP-MIB implementation
 - ◆ Linux 2.2.14
 - ◆ Implements 60 TCP-MIB Variables – including read-write.
 - » Test Applications
 - ◆ GUI interface to read and display variables
 - ◆ Prototype autotuning application

Progress To Date

- ◆ **Pre-alpha release**
 - **Halloween, 2000**
 - **Only to a few selected developers to**
 - » Test the implementation
 - » Test main concepts and TCP-MIB variables
 - ◆ See if others can develop tools that use information from the prototype.

User Support

◆ **Provided by NCSA**

- **Liaison between the user community and the developers:**
 - » Make sure Web100 software continues to address the needs of the user community.
- **In-depth support for Beta users.**
 - » Developing long-term support for larger groups of users.
- **Technical documentation**
 - » Aimed at user community.

Vender Relationships

- ◆ **Acceptance of Web100 work by OS vendors is critical to success:**
 - Cisco is already helping with funding.
 - IBM, Sun, and Microsoft are aware of the work
 - Plan to work with others
- ◆ **Web100 conference planned:**
 - Within the next 6 months.
 - Once demonstrable results are seen.
 - Vendors and Web100 collaborators.

Funding

◆ Start up funds:

- **\$100K from Cisco as part of their University Research program.**

◆ Long Term Funding:

- **NSF Grant, start date of 9/15/00**
 - » \$2.9 M for 3 years to develop the core Web100 technologies, support the early users and work with vendors.

Collaboration

◆ **Core Partners:**

- **Technical design and software development**
 - » Pittsburgh Supercomputing Center
 - » National Center for Atmospheric Research
- **User Support and GUI development**
 - » National Center for Supercomputing Applications

◆ **Early Users**

- » NGI developers and researchers to help test, debug and augment the core Web100 software suite.

Acknowledgements

- ◆ **Administrative:**
 - Wendy Huntoon (PAC)
 - Marla Meehl (NCAR)
- ◆ **Technical development Team:**
 - Matt Mathis, Jeff Semke, John Heffner (PSC)
 - Basil Irwin (NCAR)
- ◆ **GUI development**
 - John Estabrook (NCSA)
- ◆ **User Support Services**
 - Tanya Brethour, George Brett (NCSA)

For More Information

<http://www.web100.org>