Higher Speed Ethernet
An End User's Perspective

Presented by
Mike Bennett, LBNL

802.3 Working Group Interim
Knoxville, TN

Discussion

- A little background on LBNL
- LBLnet
  - Topology
  - Infrastructure
- What's wrong with LAG (802.3ad)?
- Thoughts on "motherhood and apple pie" objectives and the speed/cost formula
- Wrap-up

About LBNL

- **Founded in 1931 by Ernest Orlando Lawrence**
  - **Winner of the 1939 Nobel Prize in physics for his invention of the cyclotron**
- **Leader in science and engineering research for more than 70 years**
  - **75th anniversary this year**
- **Oldest of the U.S. Department of Energy's National Laboratories**
- **Conducts unclassified research across a wide range of scientific disciplines:**
  - **Fundamental studies of the universe**
  - **Quantitative biology**
  - **Nano-science**
  - **New energy systems and environmental solutions**
  - **Use of integrated computing as a tool for discovery**

About LBNL

- **We don’t get stock options, but we have a great view!**

LBLnet

- **LAN infrastructure and network connection support services to LBNL staff and research programs throughout the lab**
- **Since 1986, LBLnet has grown from 100 user attachments to over 13,000 device attachments, including:**
  - **Desktops and servers**
  - **High-performance clusters**
  - **Data acquisition devices and embedded systems**
  - **Building environment and security systems**
- **The LBLnet Services Group (LSG) supports a shared infrastructure in a highly heterogeneous environment.**

LBLnet Services

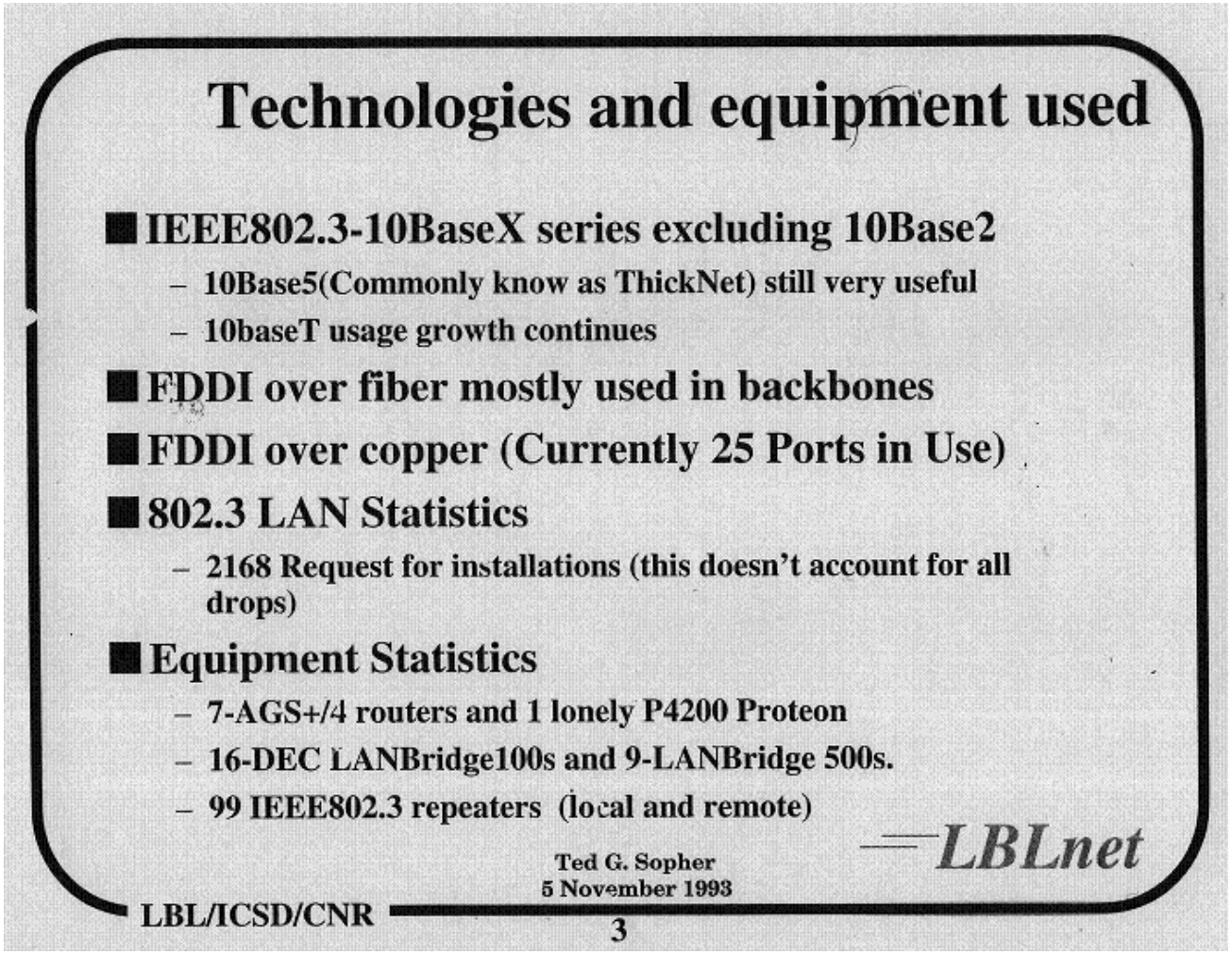
- **LSG directly administers and provides support for services including:**
  - **Domain Name System (DNS)**
  - **Windows Internet Name Service (WINS)**
  - **Boot Protocol Services (bootp)**
  - **Dynamic Host Configuration Protocol (DHCP)**
  - **Network Time Protocol Service (NTP)**
  - **Virtual Private Network Service (VPN), both software and hardware.**
- **We also work closely with the Computer Protection Program (CPP) staff**
  - **Network and systems security integrated in network at the edge and core of LBLnet**

LBLnet infrastructure

- Fibre cabling
  - 4057 strands installed
  - 1% 100/140 micron legacy MMF
  - 82% 62.5 micron MMF
    - Mix of 160 MHz * Km and 200 MHz * Km
  - 17% SMF
  - Roughly 50 buildings are connected via fibre
  - Longest route ~ 2 Km
  - Most intra-building fibre is MMF and <= 300m
  - Inter-building fibre is a mix of MMF and SMF
    - Between 200m and 2000m
  - Almost no fibre to the desktop
- Copper cabling
  - Most buildings rewired with Category 5e
  - A few places still have Category 3

Nostalgia …

- A slide from the past (before PowerPoint)



Technologies and equipment used

- IEEE802.3-10BaseX series excluding 10Base2
  - 10Base5(Commonly know as ThickNet) still very useful
  - 10baseT usage growth continues
- FDDI over fiber mostly used in backbones
- FDDI over copper (Currently 25 Ports in Use)
- 802.3 LAN Statistics
  - 2168 Request for installations (this doesn't account for all drops)
- Equipment Statistics
  - 7-AGS+/4 routers and 1 lonely P4200 Proteon
  - 16-DEC LANBridge100s and 9-LANBridge 500s.
  - 99 IEEE802.3 repeaters (local and remote)

LBLnet

Ted G. Sopher
5 November 1993

LBL/ICSD/CNR

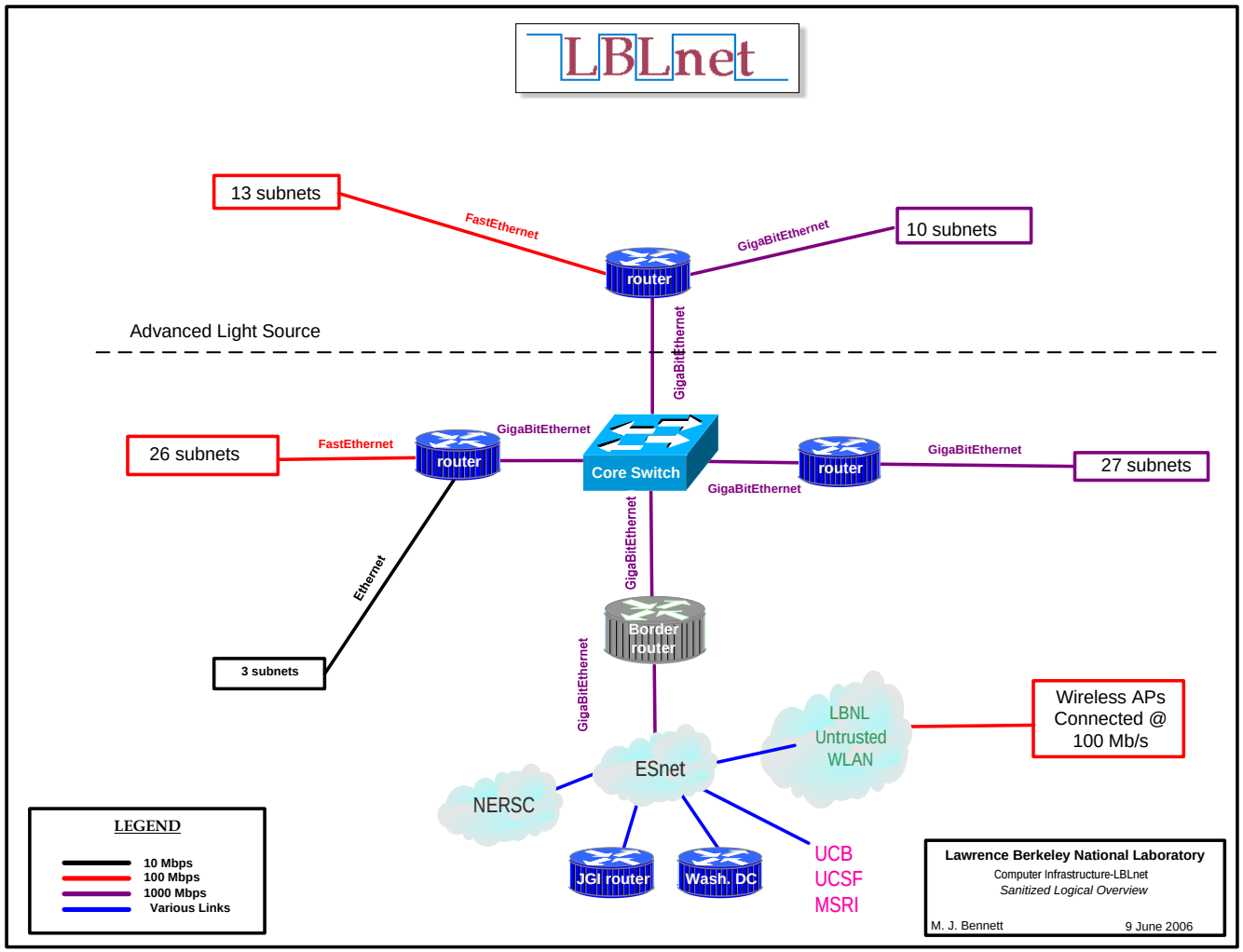
3

LBLnet technologies and equipment

- Edited to avoid perception of vendor bias …
  - 4 routers
  - 102 non-modular switches
  - 58 modular switches
  - Total: 160 switches
    - All Ethernet!
      - Shut down 2 OC3 circuits used to support microwave LAN earlier this year
- 1000BASE-SX in the core
- Of the switches mentioned
  - 69 are End Of Life (EOL) status
    - In the process of replacing them (funded project)
- 53 different versions of OS!!
  - Trying to minimize this in the EOL upgrade process

LBLnet topology

- **Simple star topology**

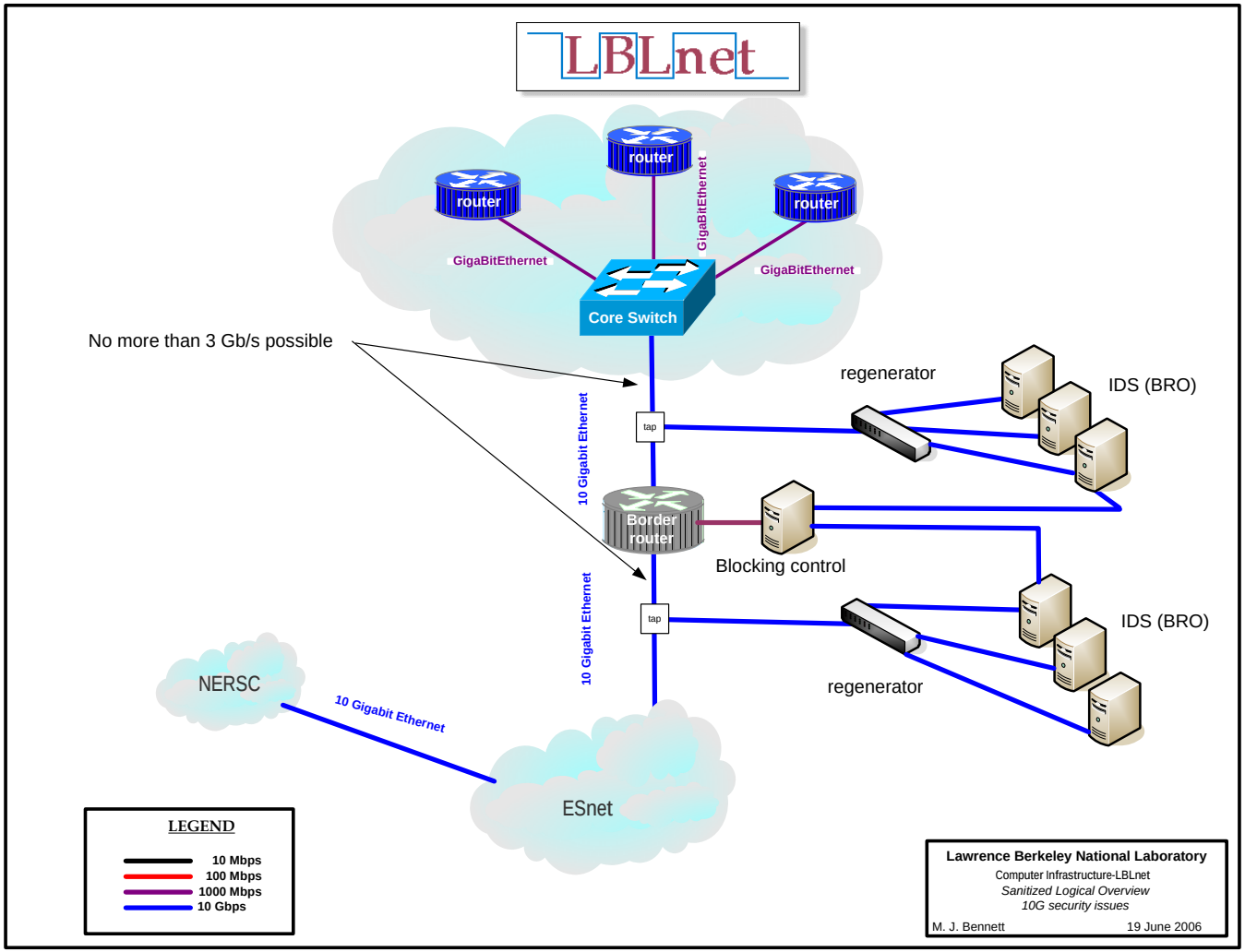


On the topic of simplicity…

- **Keeping the network simple is a prime directive for LBLnet staff. It allows us to:**
  - **Operate the network with 8 FTEs**
  - **Minimize troubleshooting and repair time**
  - **Minimize costs to the institution**
- **LBLnet Availability Report 11/1/2005 to 9/1/2006**
  - **Uptime for service switching = 99.999885%**
  - **Uptime for service VPN = 100%**
  - **Uptime for service DNS = 99.999981%**
  - **Uptime for service NTP = 99.999255%**
  - **Wireless is a secondary infrastructure**
    - **Uptime for service wireless = 99.998707%**
  - **Report excludes power outages and includes outages caused by end-users**
- **Bottom line: simplicity = lower operating costs**

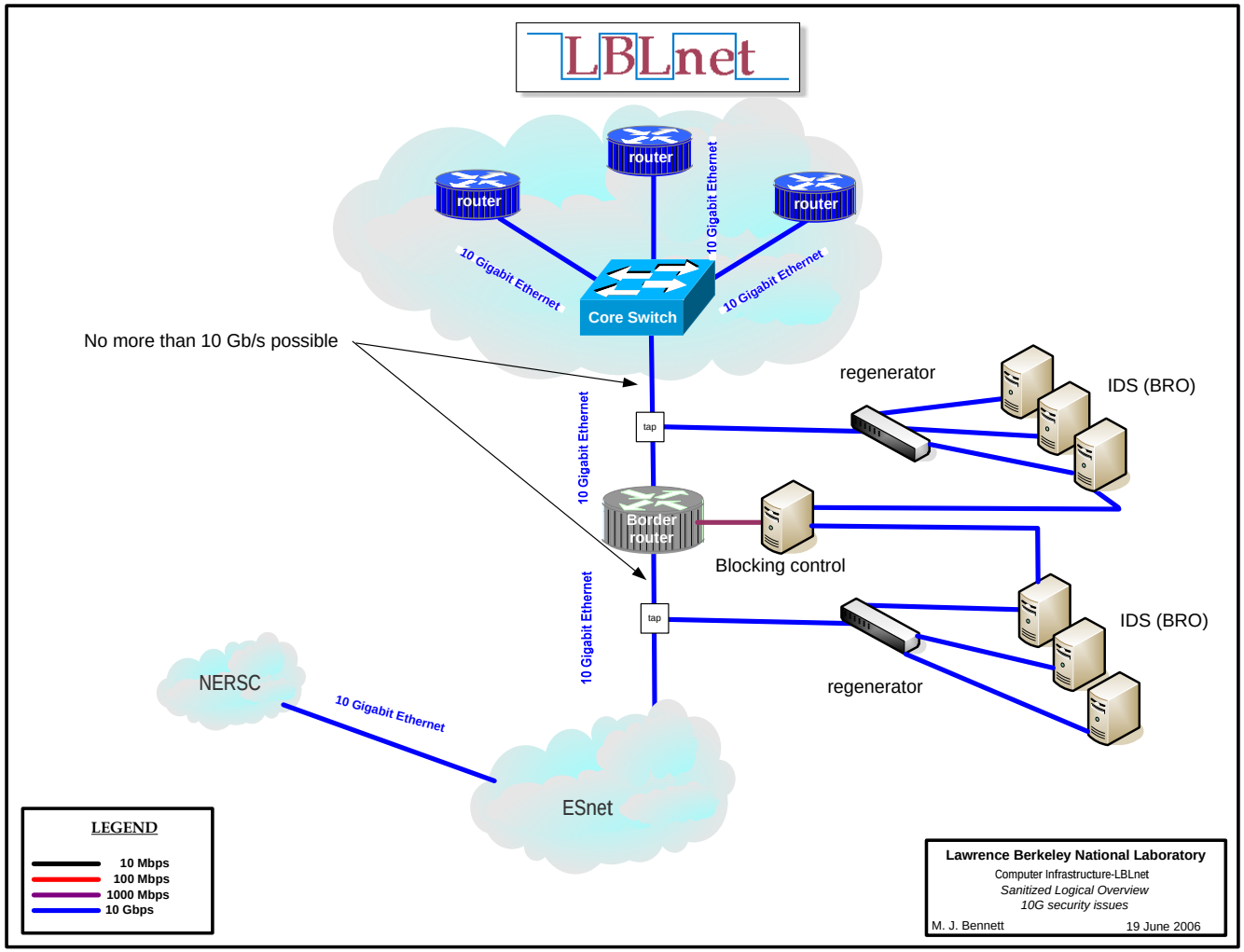
LBLnet topology

- **In a month or so …**



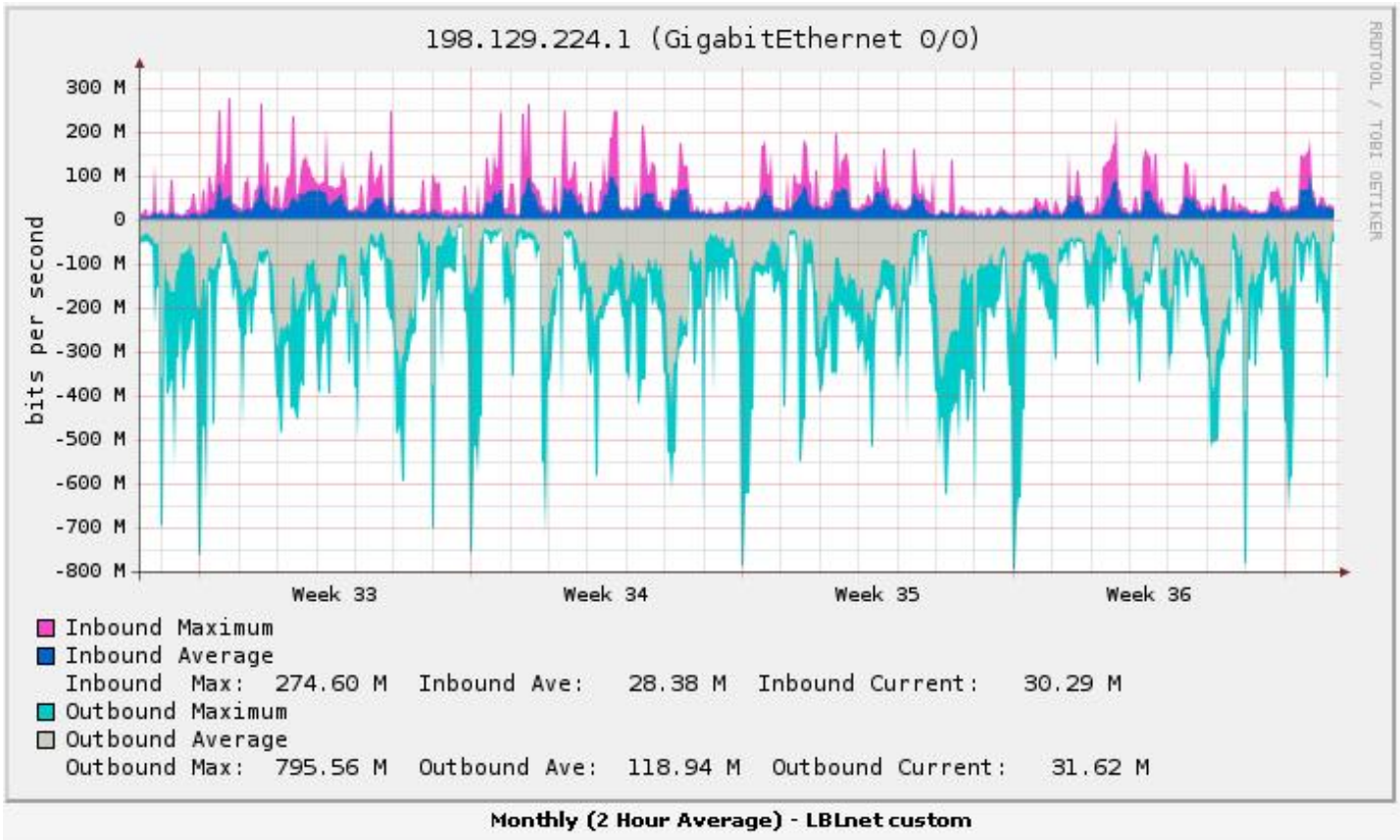
LBLnet topology

- **Next FY, assuming we get funded …**



A look at utilization

- **Monthly snapshot**



- **Note the 2-hour average**

A look at utilization

- With this utilization one may wonder:
  - Why am I a proponent of the HSSG?
    - 7 years from start of last HSSG to implementation of 10G
  - One simple answer
    - The folks who have to transport our traffic need the bandwidth
- Why not use Layer 2 Link Aggregation (802.3ad)?
  - The utilization snapshot may lead you to believe we could just add a gig port and double the bandwidth?
    - If only it were that simple …
    - Aside from dealing with IP multicast on a LAG interface, here are a few reasons why a single higher-speed link is preferred

LAG (802.3ad) - A Temporary Solution

- **Uneven distribution of traffic over LAGs**
  - **Requires vendors to implement complicated hashing functions for even traffic distribution**
  - **Requires vendors to implement per-port features on logical interface which complicates ASIC design**
  - **Most LAG/LACP implementation are limited to 8 members**
  - **Most routing protocol implementation are limited to 6 – 8 equal cost routes**
  - **Certain traffic patterns always cause inefficient link utilization**
  - **Inefficient traffic distributions for large flows**
    - **Video flows**
    - **Encapsulated traffic (e.g. MPLS, EoMPLS, IPSec Tunneling, FCIP)**
    - **>1 Gbps, 10 GbE host connections**
    - **10 GbE Transparent LAN Services (TLS) data center to data center or intra Metro services**

LAG (802.3ad) - A Temporary Solution

- LAG increases cost and complexity
- Difficult to plan for capacity and traffic engineering, especially in large networks (LAGs connected to LAGs)
- Routing optimization is best performed over fewer links
  - High bandwidth pipes are more attractive than multiple low bandwidth ones
- Higher TCO for multiple fibers in Metro/WAN deployments
  - Variable/long lead times on fiber makes it even harder to plan capacity
- Every port used for LAG cannot be used for revenue as a customer port
- Manageability/troubleshooting of multiple physical links for a single logical interface
- Difficult/expensive to implement security systems that require tapping link

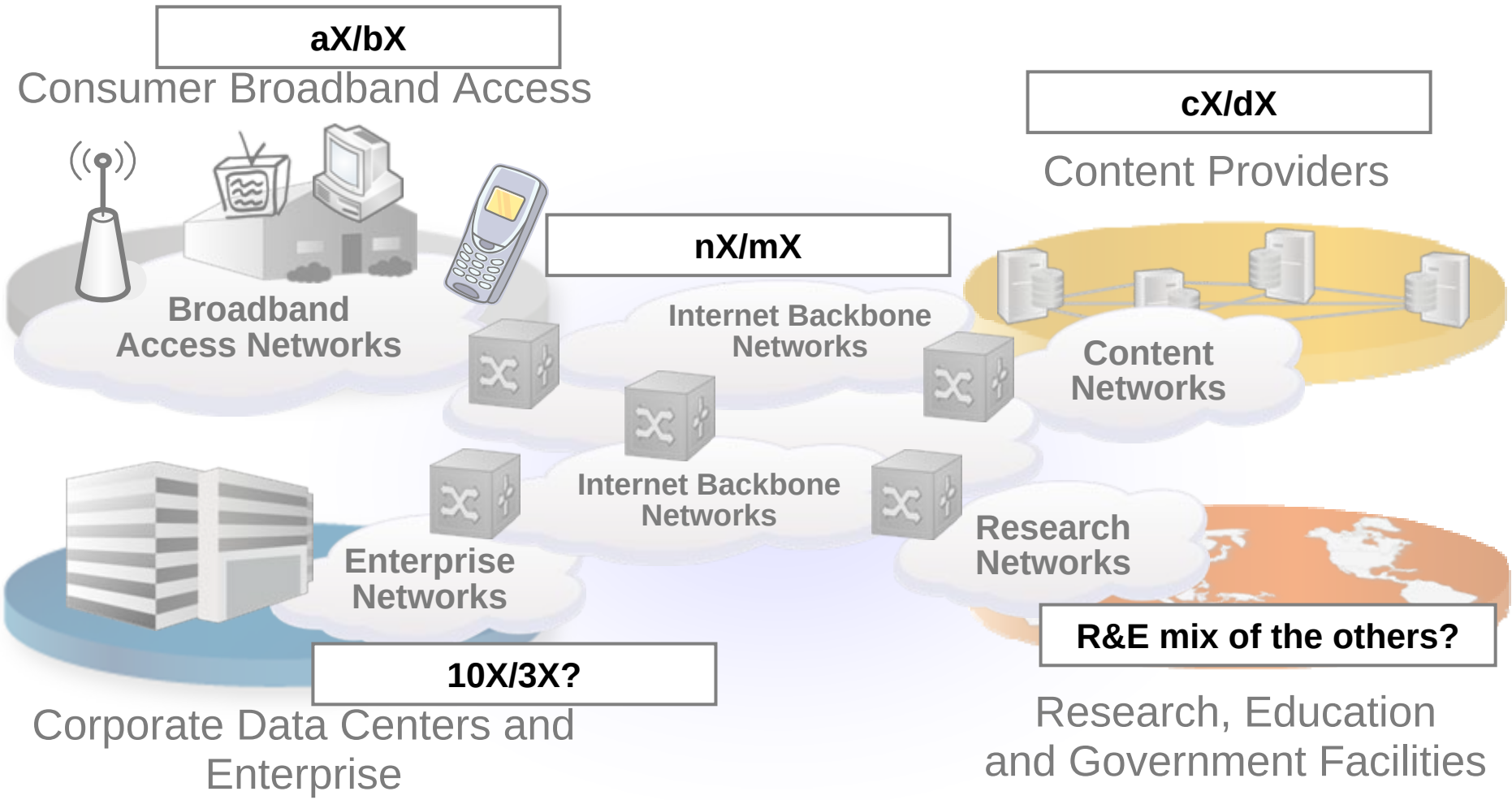
Thoughts on non-controversial objectives

- **No half-duplex necessary**
  - **Sample of 4,633 active switch ports**
    - **0 ports (1000 Mbs) in half-duplex**
    - **5 ports (10/100 Mbs) in half-duplex**
      - **Known to have hubs on the other end**
      - **Even less likely as host migrate to 1 GbE**
- **Don’t change the min and max frame sizes**
  - **No Jumbo frames necessary**
    - **We don’t use them in the LAN**
    - **I agree that we shouldn’t do anything to preclude them**
- **Need to support star-wired LANs using point-to-point links**
- **Use structured cabling referenced in ISO/IEC 11801:2002**

A thought on the speed/cost formula

- **The formula needs to be revisited**
- **One suggestion:**
  - **Create a formula per element of the Ethernet ecosystem**
    - **The LAN is where the 10X/3X formula applies**
    - **The “error” in the formula is not a matter of time to market**
    - **This would mean multiple formulae**
      - **LAN may be 10X/3X**
      - **WAN/Metro is probably higher than that**
      - **Data Center may be better**

Ecosystem speed/cost formulae



Conclusions

- **Should focus on a LAN reach objective up to 2 KM**
- **No half-duplex or changes to the minimum or maximum frame sizes**
- **Should be compatible with 802.3ad, but recognize it is a limited interim solution for increasing BW**
- **Should not get bogged down in trying to spec connectors, etc**
  - **No connector wars!**
- **We need to redefine the speed/cost formula**
  - **Perhaps define a few**

Thanks!

- **Questions or comments?**
  - **mjbennett@lbl.gov**