A User's View of Gigabit Ethernet

Bob Fink

Lawrence Berkeley National Laboratory

IEEE 802.3z meeting

Coeur d'Alene, Idaho

9-11 September 1996

Time for another network upgrade

- **Our Lab net has been in existence since mid-1983 and numerous upgrades have been done over the intervening 13 years:**
  - **First fiber plant (100 micron) – 83-84**
  - **Bridging – 85**
  - **Second fiber plant (62.5 micron) – 87-89**
  - **Structured Cat 3 wiring – 87-89**
  - **FDDI Backbone – 88-89**
  - **Routing – 89-90**
  - **Twisted-pair Ethernet – 89-90**
  - **LAN switching in selected “hot spots” – 90-91**
  - **FDDI in servers and very few workstations – 91-92**
  - **numerous upgrades to bridges and routers – always !**
  - **Fast Ethernet and LAN Switching in backbone and on workstations – 95-96...and it continues**

the Problem

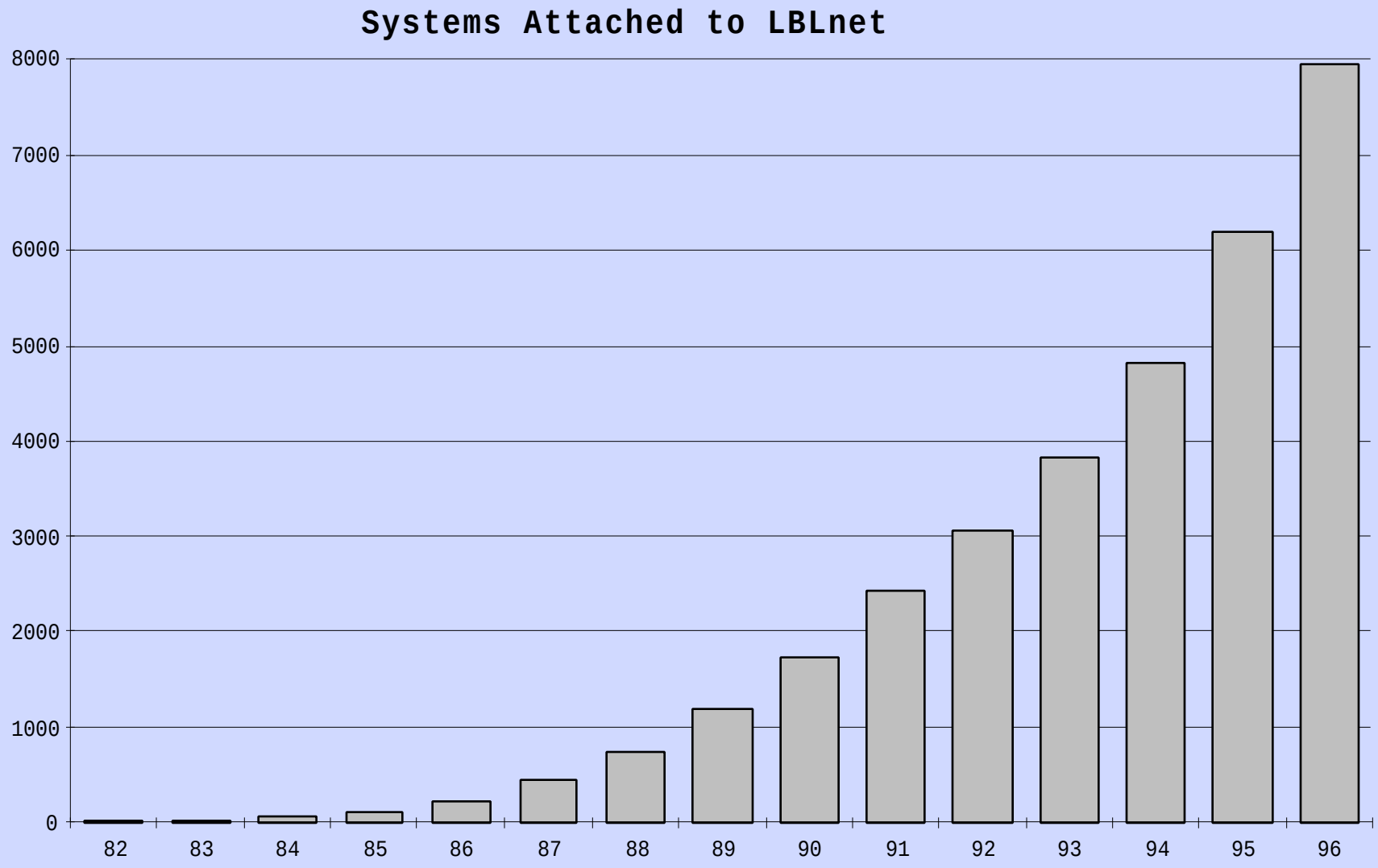
the Berkeley Lab

- **mostly one of performance and growth in network capable systems:**
  - **load problems caused by shared LAN media use**
  - **users needing higher performance links**
  - **routers/bridges/switches needing higher performance links**

    and
  - **the right technology to appear at the right cost!**

Workstation growth at LBNL

the Berkeley Lab



Next Generation Systems

- **The National Energy Research Supercomputing Center has just moved to LBNL**
- **A Cray T3E has just been installed, so for at least a month we are the largest supercomputing center in the world :-)**
  - **This puppy has 5 ‘gigarings” each capable of 400 GB/sec full duplex for I/O - collectively that’s 4 terabit/sec of I/O**
  - **And believe it or not it is still connecting to the outer world with half duplex FDDI !**

Nets of Workstations

- having said that to impress you, we really believe the future is in high performance networks of workstations (clusters if you will)
- Major performance gains in memory bandwidth have been made in just the last year – this is one of our key indicators of overall system performance
- For typical medium level workstation – $10-20K
  - One year ago: 20 - 25 M Bytes/Sec
  - Now: 130-200 M Bytes/Sec

What they need for networking

- high performance low cost LAN switching
  >10 Gbps switching capacity

- high performance low cost full duplex links
  1 Gbps to start

- graceful scaling,
  i.e., 10/100/1000
  (no nasty translations or transitions)

Full duplex good

- **Shared media LANs such as Ethernet and Token Ring (including FDDI) were designed as half duplex**
- **Thus the performance is half what it might be**

  (you don't have to believe in this absolutely, but we slowly get there by beating on w/s designers for I/O efficiency - Van Jacobson et al)

- **The answer – Full Duplex and LAN Switching of course**

Fast Ethernet

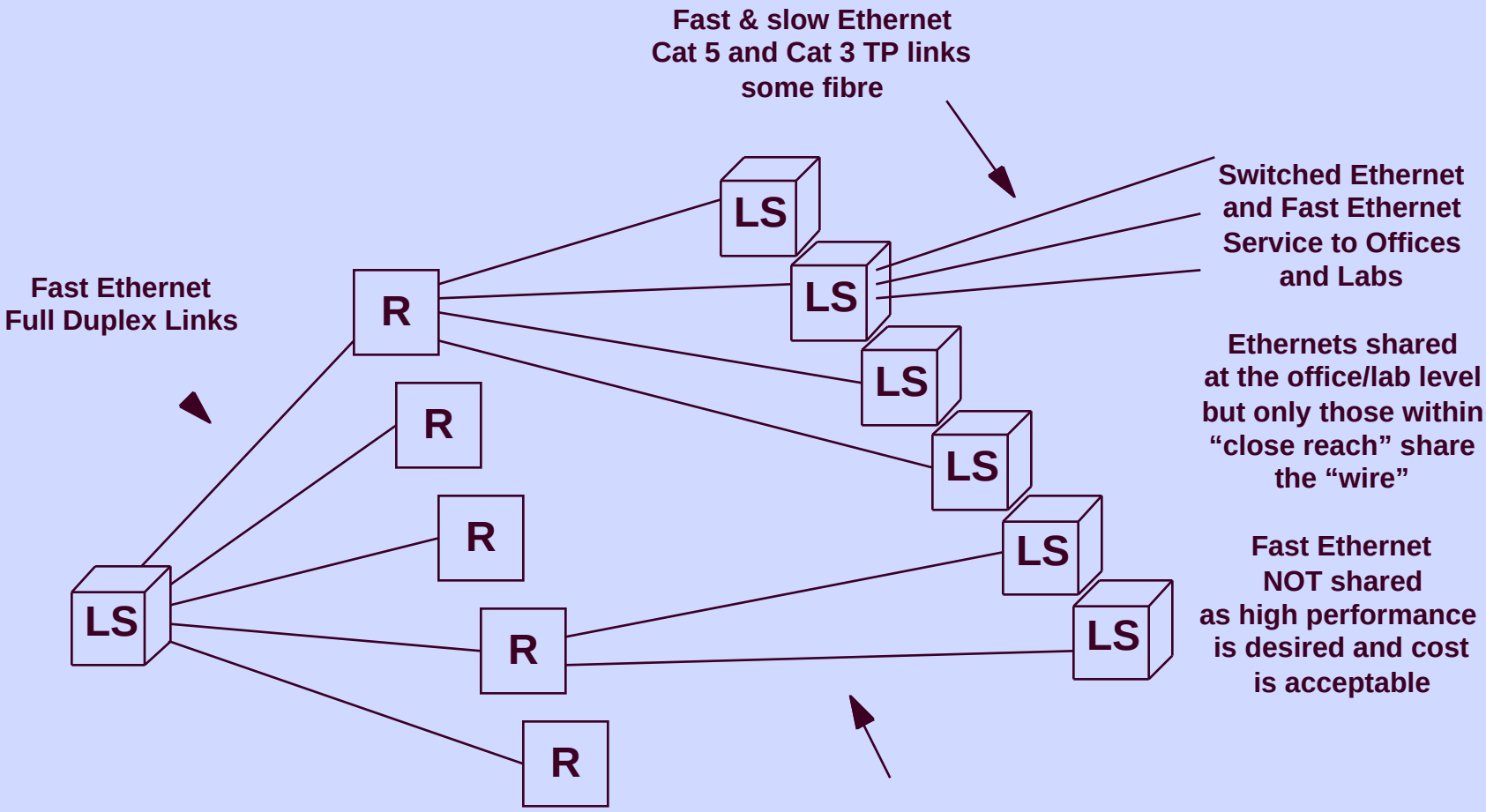
- Can operate in an ~200m shared bus environment
- What is really nice about this is NOT that it can be shared, but that it uses our structured wiring plan of up to 100 m Category 5 TP, so we could delay the inevitable conversion to fibre for a few years
- cost effective LAN switches for FE allowed users to choose to run single system “wires” thus utilizing FE’s full duplex feature - this is the only way to date our users have chosen to use FE!
- ... and it is totally compatible with Ethernet

LAN Switch Issues

- LAN switches, like their ATM cousins, are only as good as the buffering and queuing strategies they employ (among other things)
- it is only recently that memory, cpu-s and hybrid LAN controller logic have become cheap and fast enough to provide high performance, low price LAN switches
- even then, many are junk, while others are brilliantly executed – and it keeps getting better
- deep LAN market penetration of Fast Ethernet has completed the equation with a high performance link to routers and users

Our Overall Architecture

the Berkeley Lab



Progress in the last 12 months

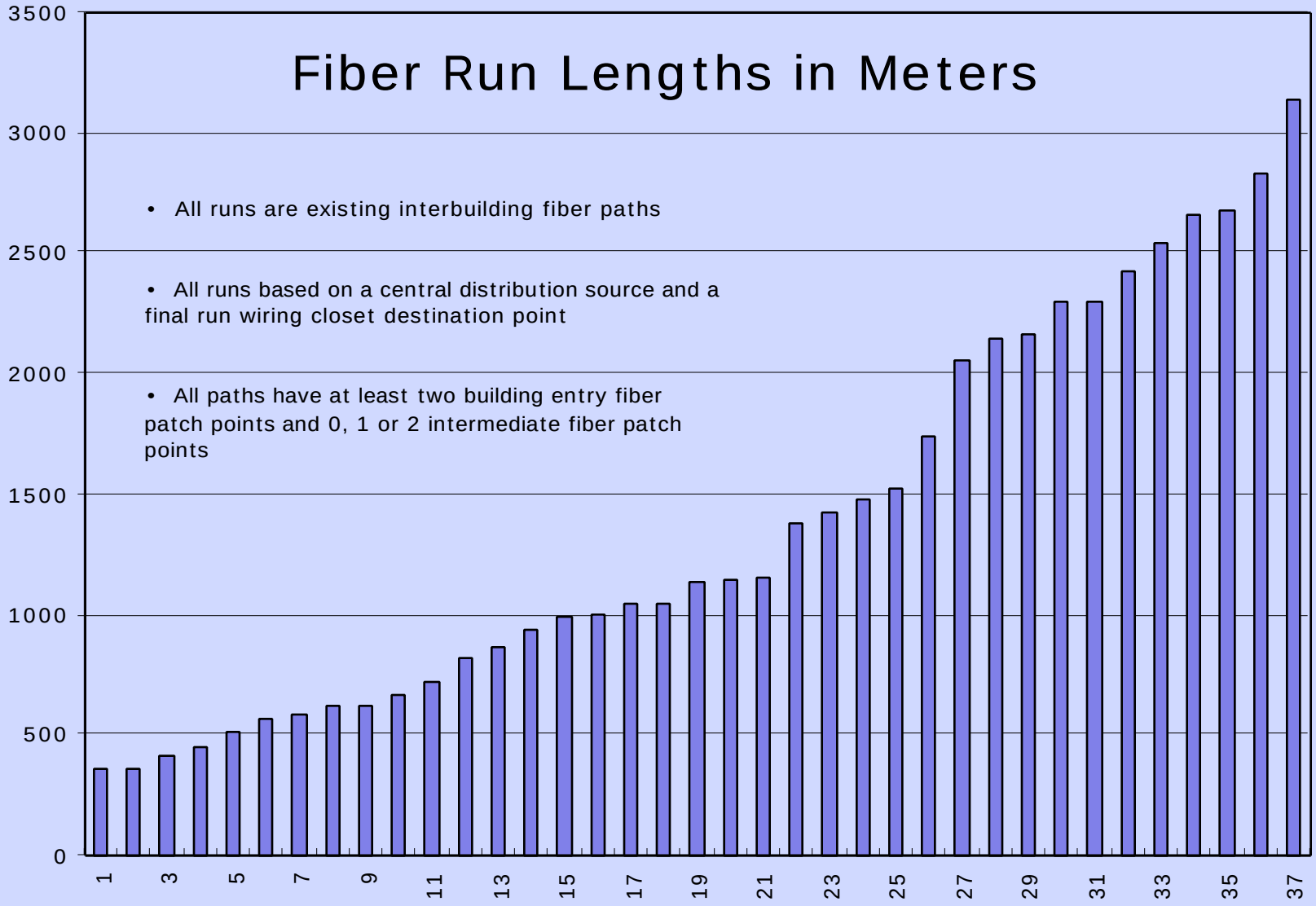
- **Replaced all our old routers with new high-performance ones capable of supporting lots of FE**
- **Almost completely phased out FDDI (only legacy systems left)**
- **Bought 1200 switched Ethernet ports and 500 FE ports**
  - **The Ether ports will be loaded in the range of 1- 5 devices per port, depending on need**
  - **The FE ports will be loaded at an average of 1 per port and be run full duplex**

Fibre optic links

- **Fibre links used inter building, with distances from 350 m to 3000 m (splice/connector loss and modal bandwidth tightly controlled)**
- **Some national labs have run FDDI MM PMD links in the 5 - 10 km range (good loss control!)**
- **Multi mode used till now for Ether, FDDI and Fast Ether**
- **Single mode being installed for runs longer than 500 m**
  - **a $250K fibre enhancement project now underway has been adapted to SM requirements of Gigabit Ether for our longer inter-building links**

Fibre Length Distribution

the Berkeley Lab



Office Drops

- Very few fibre office drops (i.e., closet to office final 100 m runs) have been installed to date
- Site wired for Category 3 structured wiring in 1987-89, all new drops are Category 5 since 1994
- Costs for new Cat 5 double 4-pair office drops runs in $200-300 range, while two MM fibre office drops cost in the $600-800 range
- Cost of new Cat 5 or fibre office drops have NOT been a deterrent to users switching to FE, no reason to expect this will be less so with GE

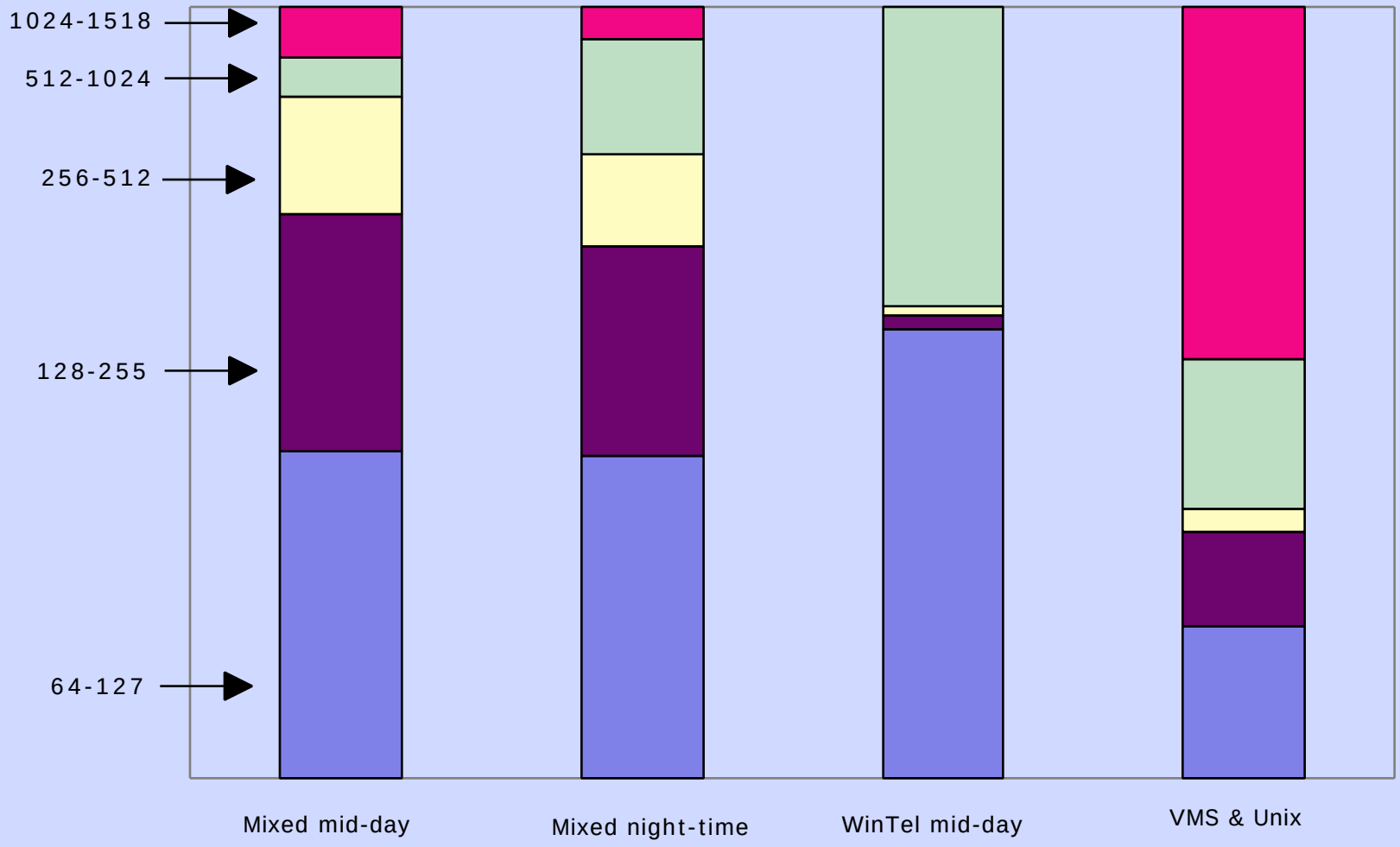
FE & LAN Switch Experience

- **No failures to date**
- **No performance problems to date**
- **Stress testing to date has not uncovered problems**
- **Heavy efforts are now beginning to manage these nets and devices**

lots of issues around this - but not for here!

Sample Packet Distribution

the Berkeley Lab



Conclusions on packet size

- Can't conclude anything as there is so much variance per subnet and over time
- When looking at the "workgroup average" packet size used for various 802.3z simulations, one can conclude that its sizes are in the range of our highly varying distribution...
- ...but I would hesitate to ever draw any conclusions based on ANY view of "average" size distribution

So what does this mean for Gigabit Ethernet?...well, IMHO

the Berkeley Lab

- **LW laser choice for use over MM and SM seem good**
- **I don't believe efforts on "half duplex" CSMA/CD operation to be worth the price and added complexity:**
  - **our users are almost totally going for FE switching to date (the one exception was our first FE user, and he is now going to fully switched FDX as well)**
  - **initial use for GE will be in backbone/feeder nets while switching prices come down**
- **Having said that, if you still must do it, the 200 m minimum diameter is essential, less won't do**

...more

- **Concern about the Mac level reliance on signaling hidden from the user's packet level as it impacts passive monitoring ...**
  - **the more features to hack around at the handler level, the more costly passive monitoring is...**
  - **at least try to include ways to turn it off**
- **I believe that reliance on past work is essential, i.e., use of FC ideas and components is great**
- **KISS - everything done to keep the design simple will pay off in the long term**

That's all folks

- **Thanks for the opportunity**
- **Clearly these kinds of collaborative efforts are working better than ever**
- **Keep the focus, stay on track, and forge on...**

(Hope I get to work on 10 gigabit Ether before I retire - any bets on how long before 1518 byte packets are too small? ;-)