DAG Network Measurement

Dr Stephen Donnelly

Endace Technology

<stephen@endace.com>

Origins

- The University of Waikato
  - Professor Ian Graham
- TeleSim
  - Canadian Research Project
  - Large Parallel ATM Network Simulation
- How to verify Simulations?
  - Traffic Models
  - Router Models

Approach

- Build Custom Hardware
- Unusual Interfaces – OC3 ATM
- Full line rate performance
  - Worst case
  - Simulation
  - Attacks
- Time stamp quality

Early Work

- DAG
  - Daughtercard for commercial NIC
  - DS3 and OC3
  - ATM Headers only
- DAG 2
  - PCI card
  - OC3 ATM
  - Full cell capture

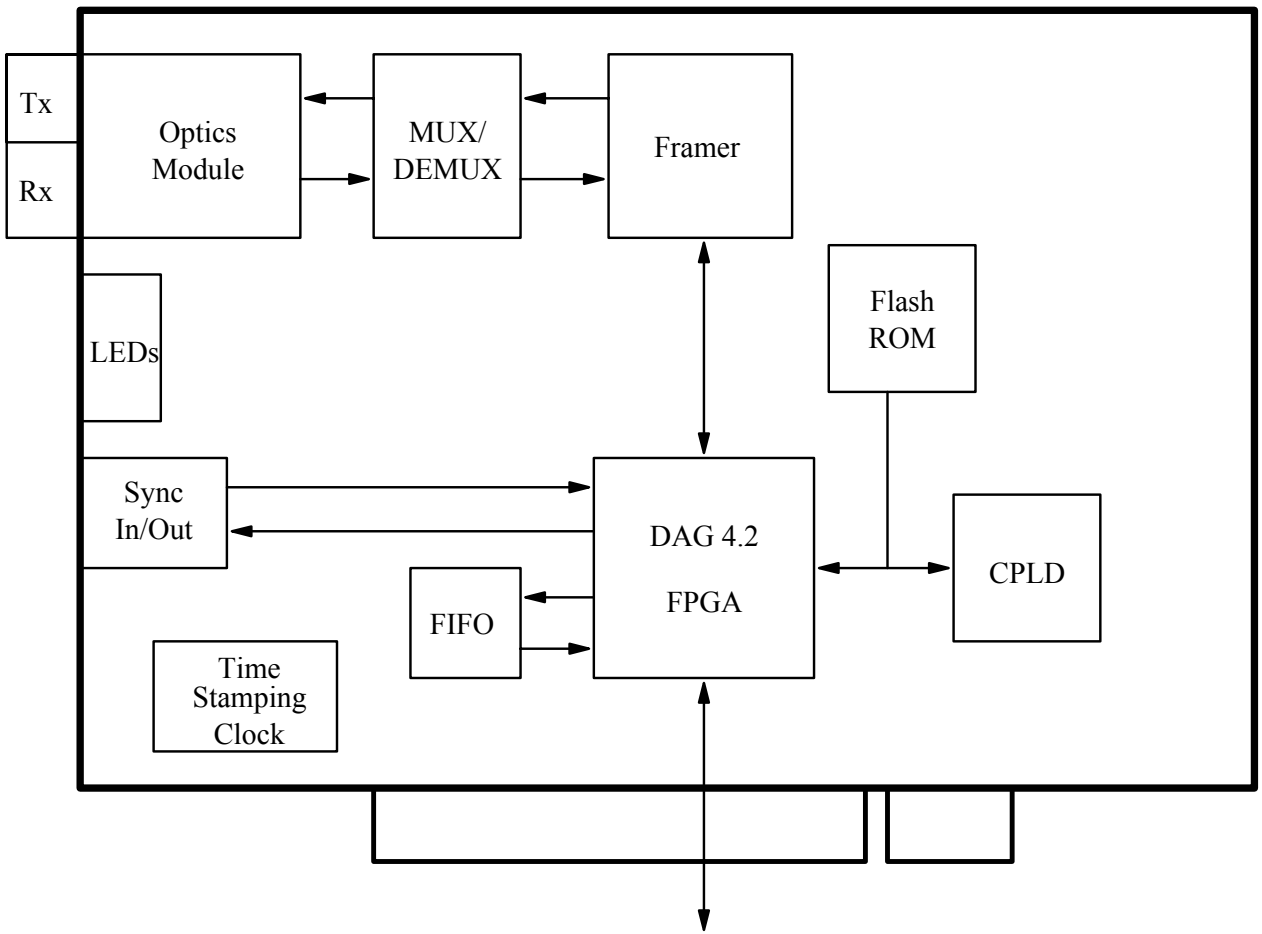
Collaborations

- DAG 3
  - NLANR PMA
  - Sprint ATL “IPMON”
  - OC3 and OC12 ATM and POS
- DAG 4
  - CAIDA “OC48MON”
  - OC48 ATM and POS

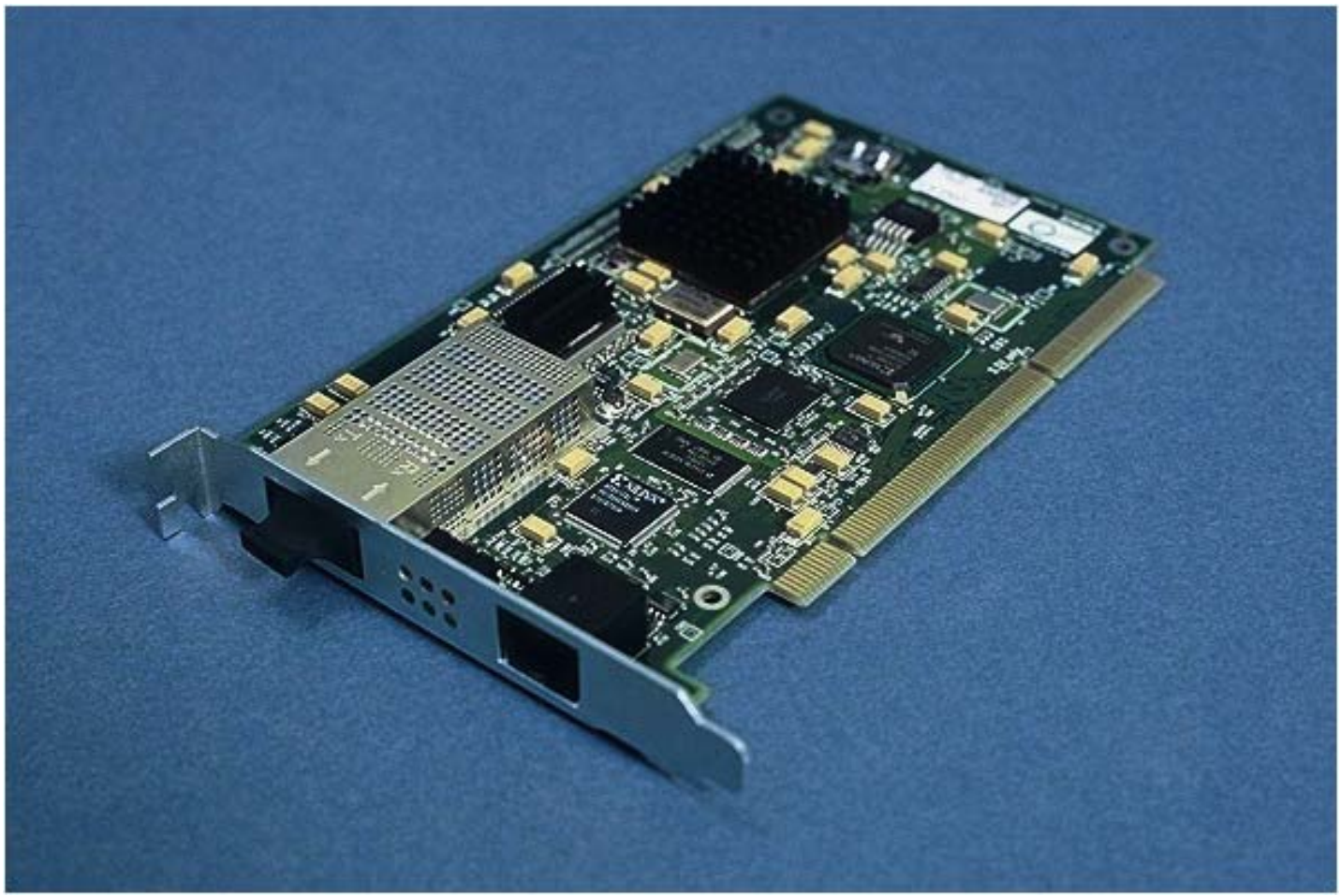
Endace

- Research Group
  - Several International Customers
  - $1.4M turnover
- Spin off companies
  - EMS Production and Sales
  - ET Design and Development

DAG 4 Architecture



DAG 4.2



DAG 4.2

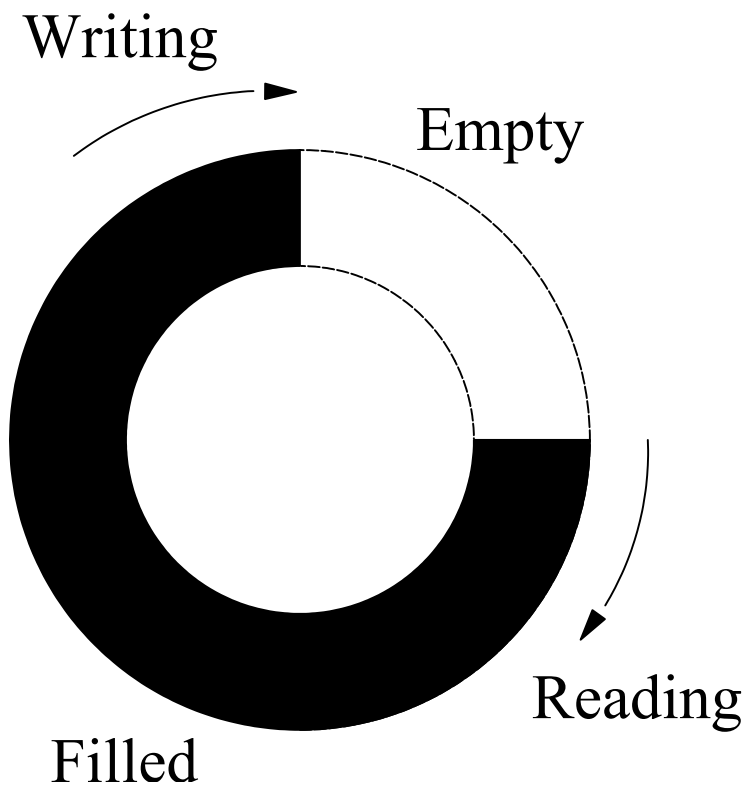
- OC48 2.48832 Gbps
  - Minus SONET overhead; 2.39616 Gbps
  - 299.52 MBps
- PCI 2.2
  - 64-bit 66MHz
  - Theoretical Peak 528 MBps
  - Actual Efficiency ~60-70%
  - Realised ~350MBps

Software Pitfalls

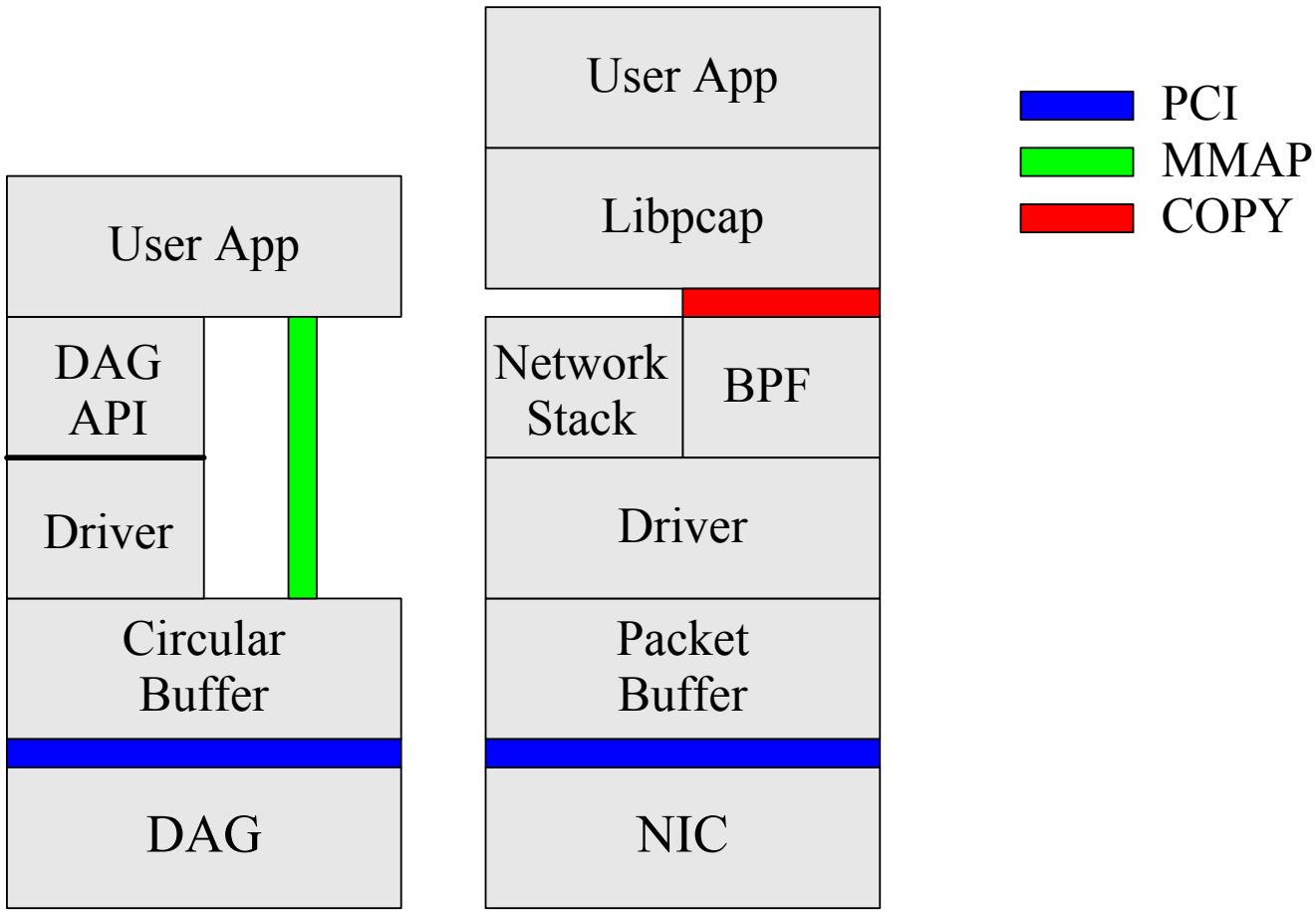
- Potential Traps
  - Avoid Per-packet Overhead
  - Avoid Dynamic Buffer Memory Management
  - Avoid Network Stack
  - Avoid High-rate Interrupts
  - Avoid Copying
- Answer?
  - Capture process must be light-weight

DAG Software Approach

- Large Static Circular Buffer
  - Avoid alloc/dealloc
- Window based handshaking
  - Not per-packet
- Memory mapped IO to User-space
  - Zero-copy



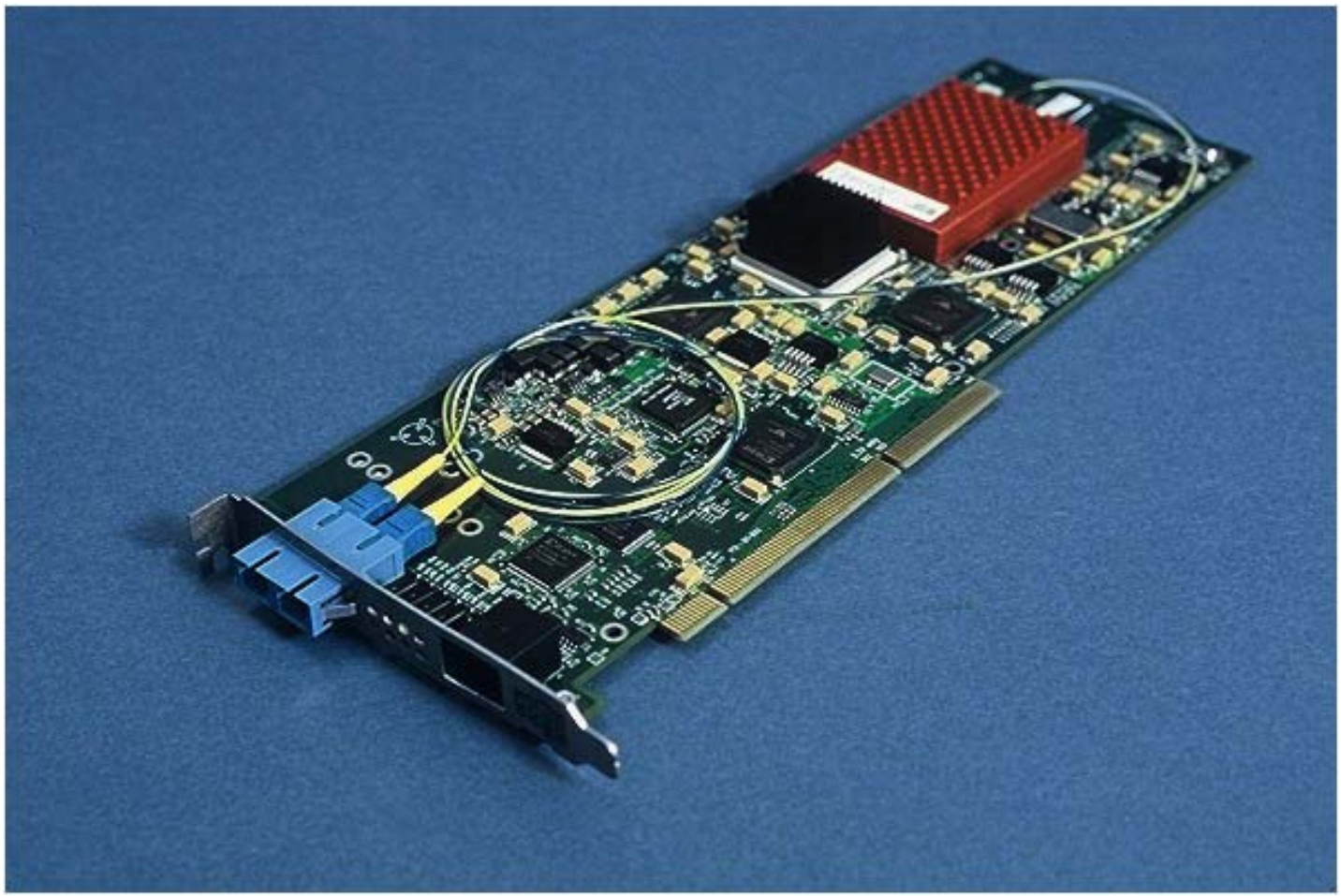
Software Layering



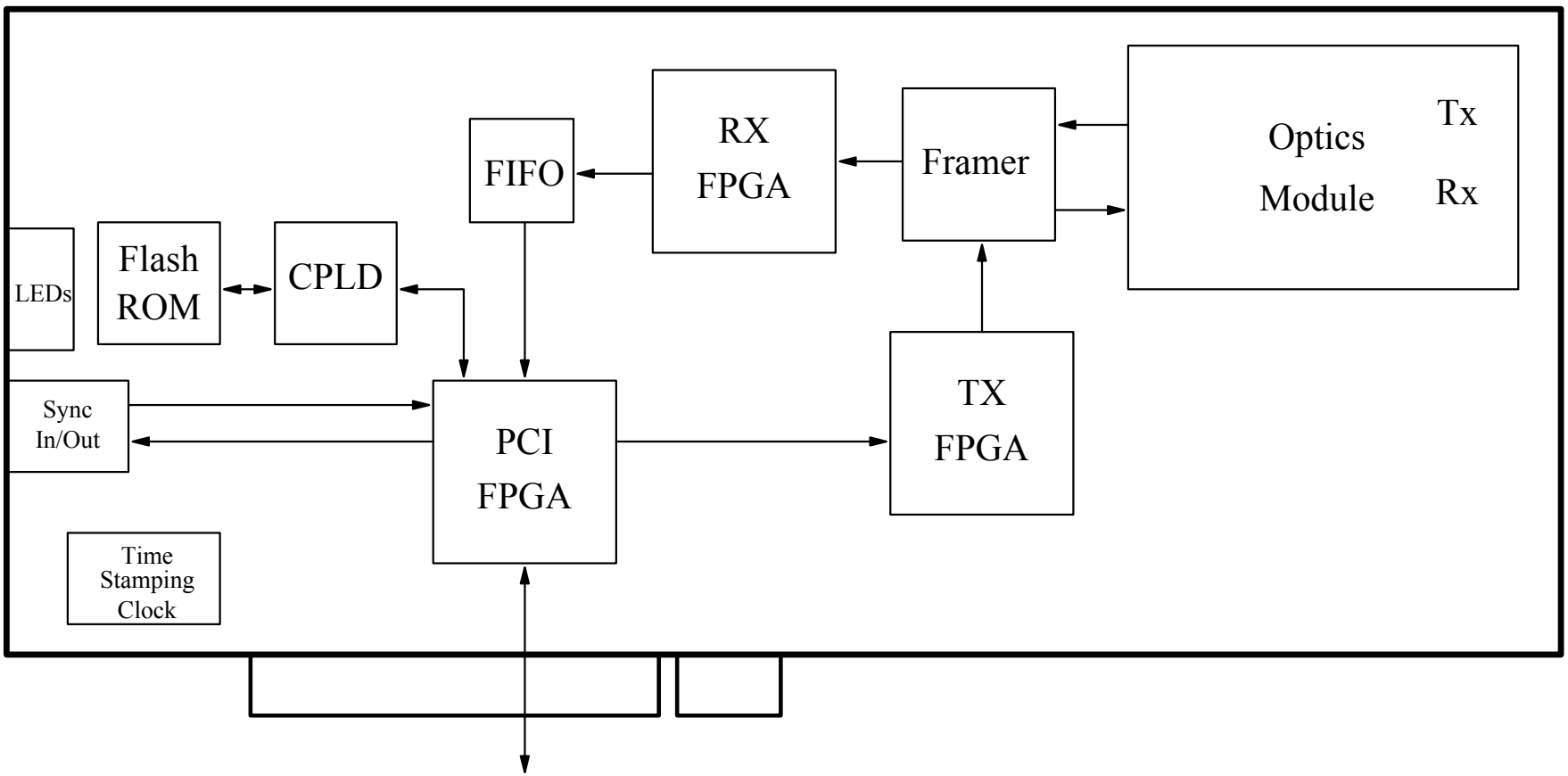
Actual Performance

- Single PC
  - 2U size
  - Dual processor Xeon 1.8GHz
  - Two DAG 4.2 OC48 cards
- Capture two OC48 links at full load
  - 5Gbps
  - 12Mpps
  - Delivered continuously to userspace

OC192 – DAG 6.0



DAG 6.0 Architecture



DAG 6.0

- OC192/10GE prototype
  - OC-192c / STM-64c POS, 10GBase-W Ethernet
  - Intended to test network interface electronics
- DAG 4 like back-end
  - PCI 2.2 = ~350MB/s
  - 10Gbps = 1250MB/s
  - Less than full load

Field Trials

- Successful field trials in 2002
  - USA
  - France
- Operational OC-192c POS Backbone Link
  - 25% load average = 2.5Gbps, 0.5Mpps
  - 600 Byte average size, capture 48 Bytes
  - Resulting data stream of 32 MBps
  - Easily written to Hard Disk

DAG 6.1

- Production OC-192 card
- PCI-X 1.0
  - 10Gbps = 1250MB/s
  - 64-bit 133MHz = 1064MB/s = 85% load
  - Efficiency ~80% -> 70% load
- Limited by available bus technology
  - PCI-X 2.0 Coming -> 266MHz, 533MHz
  - Motherboards first, FPGA limits

DAG 6.1 Outlook

- Full packet capture to ~70% link load
- Packet header capture to 100% load
  - Average packet size 200-600 Bytes
  - TCP/IP header size 40 Bytes typ.
  - 5:1 to 15:1 reduction
  - Worst case ~24Mpps (<40 Bytes Avg)
    - High ratio of overhead to data inflates load
  - Field trial 25% load = 0.5Mpps

Too Much Data?

- We can capture a lot of data, but can we process it?
  - Increasing link rates mean more raw data
  - Extracting information requires more work
- Reduce incoming data volume
  - Sampling – Trivial
  - Partial Capture – packet headers
  - Filtering – capture only select packets
- Hardware Application Acceleration

Filtering

- Accept/Reject
  - Reduces data volume to CPU
  - Don’t know about rejected traffic
- Packet Marking/Colourisation
  - Classification/Pre-sorting
  - Extract Parallelism
  - Does not reduce volume to CPU

Filter Types

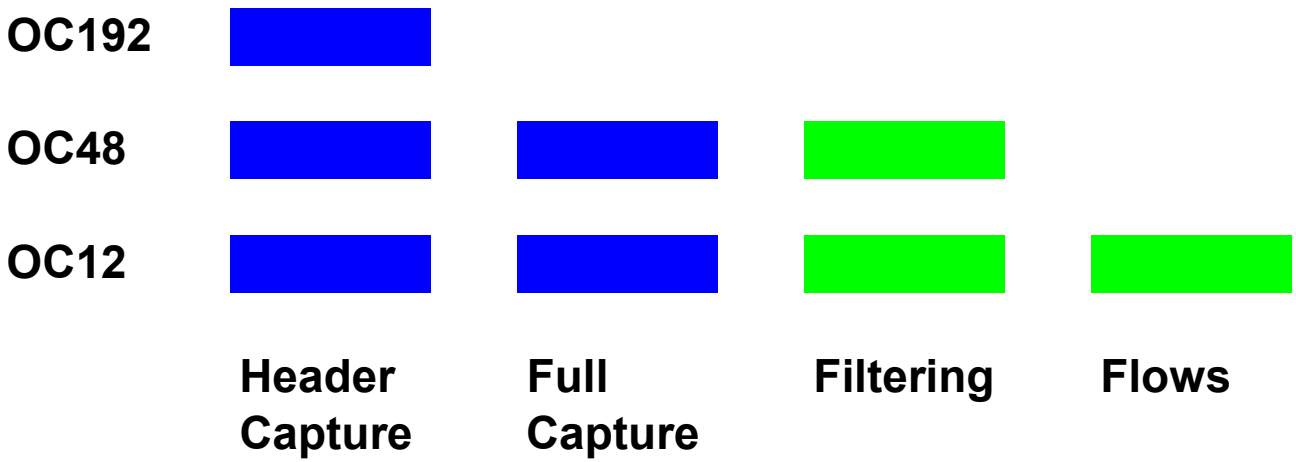
- Address related filters
  - Netblocks
  - Longest Prefix Match – AS Lookup
- Generic Filters
  - Operate on any IP/TCP field
- How many are useful?
- MPLS and variable length link headers
- Endace design for OC48, what about 192?

Flow Analysis

- Traffic Eng, Accounting, QoS, Routing
- Hardware Flows Acceleration
  - Flow colourisation
  - Full Flow Export
- IETF
  - RTFM
  - IPFIX

DAG 3.8 Dual OC3/12

Future



- What does SCAMPI need?
- We want to work with you!

Links

- http://www.endace.com
- http://dag.cs.waikato.ac.nz
- http://wand.cs.waikato.ac.nz
- http://pma.nlanr.net
- http://www.caida.org
- http://ipmon.sprintlabs.com
- http://www.pcisig.com