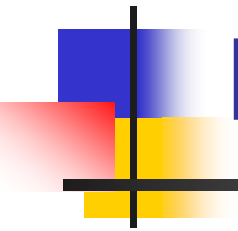
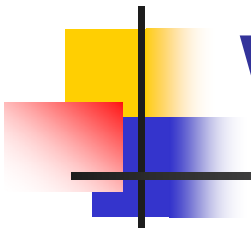


DragonFlyBSD

Jeffrey Hsu

Member of FreeBSD and DragonFlyBSD Project
hsu@freebsd.org
hsu@dragonflybsd.org

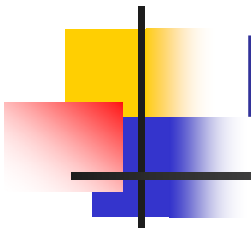


What is DragonFlyBSD?

- FreeBSD
- People
- Organization
- Today and tomorrow’s technology

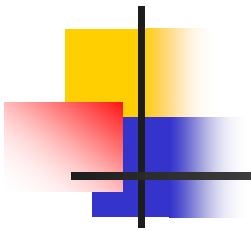
Project Goals

- Fast
- Stable
- Suitable technology
- Efficiencies of scale



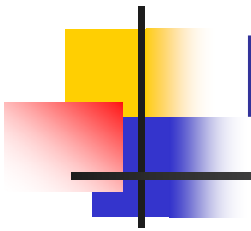
New Concepts

- Messaging as kernel structuring mechanism
- Application-specific customization
- PMP
  - NUMA, not SMP
  - Contention-free parallel multiprocessing



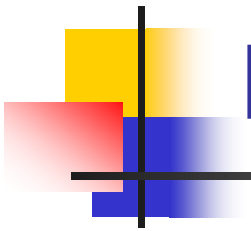
Infrastructure

- Lightweight threads
  - User and kernel-land
- Messaging
  - Predicate messages
- Tokens



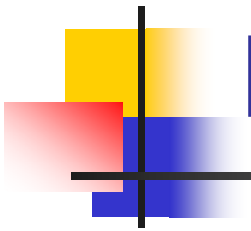
Lightweight Threads

- Separate execution context from address space
- Guarantees for performance
  - Allows per-cpu data with locking
- Separate LWKT scheduler
- User-land messaging



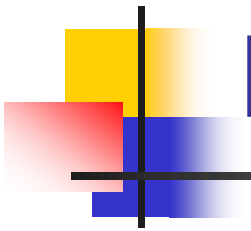
Network Stack

- Extensive use of messaging
- Protocol enhancements
- Implementation improvements



Network delays

- 50ms RTT worse than 10ms disk seek time
- 4 RTTs even worse than 1 RTT
- 1 second min RTO even worse than 4 RTTs
- More than half of retransmits are timeouts
  56% timeout
  44% Fast Retransmit

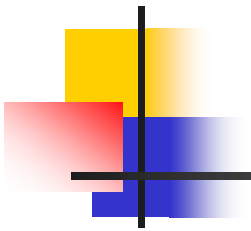


Protocol Enhancements

- NewReno
- Larger initial window size (RFC2414)
- Limited Transmit (RFC3042)
- Eifel detection (RFC3522)
- Early Retransmit
- More good stuff coming!

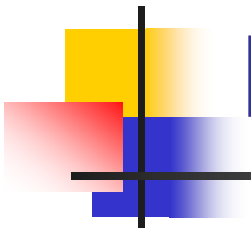
Limited Transmit

- On each dupack, send out new data
- helps with small send windows
- congestion window of 3
- conservation of packets
- 44% Fast Retransmit
  56% timeout
    4% saved by SACK
    25% saved by Limited Transmit



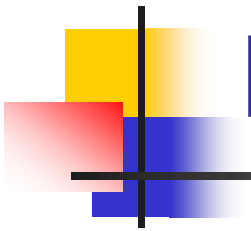
Implementation improvements

- Parallel MP design rather than serialized SMP design
- Costs of networking
  - TCP syncache
  - UDP transmission
- Hardware offload for TCP segmentation



Hardware TCP Segmentation

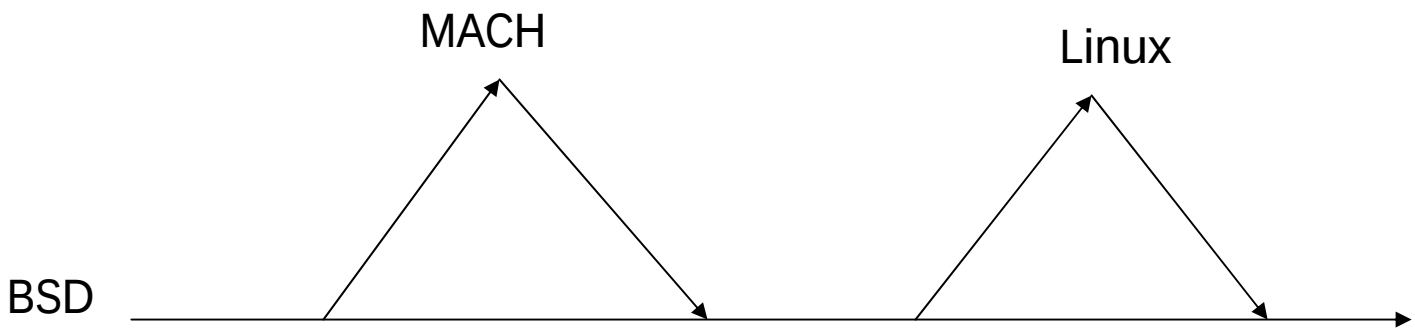
- Send large packet down to NIC.
  NIC breaks it up and send out lots of MSS-sized packets.
- Savings
  - CPU cycles on host to do segmentation and going down the network stack several times
  - Minimize I/O bus crossings
  - One large DMA setup and transfer
  - Transmit complete interrupts

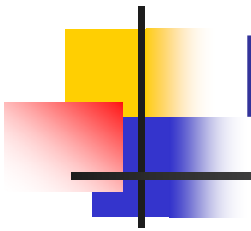


MP support

- Partitioning and replication
- Lock-free multi-processing
- Early packet classification
- Cohort scheduling
- Generic framework
- Incremental deployment

Historical perspective

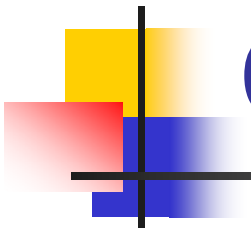




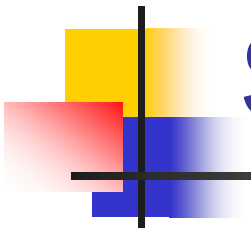
Predict Future

- Lots of ideas to explore
- Lots of projects to maintain

Call for Involvement



- http://www.dragonflybsd.org
- news://nntp.dragonflybsd.org
  kernel, submit, commit, bugs
- http://www.freebsd.org/~hsu/papers/dragonflybsd.asiabsdcon.pdf
- Submit bug patches, code, project ideas
- Maintain subsystems: acpi, pc98, IPv6



Summary

- New BSD with exciting possibilities
- People
- Interesting technical trends
- Stability and performance
- Call for involvement