Network Subsystems Reloaded: A High-Performance, Defensible Network Subsystem

Anshumal Sinha, Sandeep Sarat, Jonathan S. Shapiro

Systems Research Laboratory
The Johns Hopkins University
Monolithic Network Subsystems

- Issues with the monolithic network subsystems
  - Single point of failure.
  - Difficult to achieve per-client resource accountability.
  - Robustness-critical code is large.
  - Difficult to maintain and debug the code.
  - Simultaneous existence of multiple protocols not easy.
  - Limit the ability to perform application specific optimizations.
Why Use Monolithic Network Subsystem

- Performance !
- Performance !!
- Performance !!!

- No viable alternative.
Background

- Previous attempts:
  - Thekkath et al. [93] implemented user-level network on Mach 3.0
  - U-net [95] implemented user-level network for parallel and distributed architecture.
  - Exo-kernel [02] provided application level resource management in Xok/ExOS.

- Failed to deliver sufficient throughput using commodity hardware and address defensibility at the same time.
Hypothesis

- Key issue in factoring is avoiding data motion.
- Earlier operating systems did not provide appropriate properties to facilitate such factoring.
- EROS supports domain factored design.
Methodology

- Built two network systems:
  - EROS Monolithic Network Subsystem
  - EROS Domain Factored Network Subsystem
- Both subsystems based on lwIP stack.
  - Easy to port.
  - Not optimized for performance.
- Drivers based on Linux.
EROS Monolithic Network Subsystem

- Similar to monolithic network subsystem.

![Diagram showing the relationship between Hardware, Kernel, User Level, Monolithic Stack, Client Application, T: Timeout Helper, I: IRQ Helper]
EROS Domain Factored Network

- **Data**
- **Headers**
- **Client Application**
- **Network Stack**
- **Enet Layer**
- **Shared Region**

**T**: Timeout Helper
**R**: Receive Helper
**I**: IRQ Helper

**User Level**
**Kernel**

**Hardware**
Various Network Subsystems

- Traditional Monolithic (Linux)
  - Client Application
  - N/W Stack
  - Ethernet Driver
  - Hardware

- EROS Monolithic
  - Client Application
  - N/W Stack
  - Ethernet Driver
  - Hardware

- EROS Domain Factored
  - Client Application
  - N/W Stack
  - Ethernet Driver
  - Hardware
Performance:

- Measures of interest are latency and throughput.
- Ethernet Cards 100Mbps and Gigabit Ethernet.
Conclusion

- Factoring is more feasible than previously assumed.
- Supports Blackwell’s claim on significance of instruction cache locality.
- Factoring provides basis for building defensible system.
The End.
Performance: ttcp

- ttcp
  - 100M : 32768 bytes
    - Linux : 11.54 MBps
    - EROS Monolithic : 11.62 MBps
    - EROS Factored : 11.61 MBps
  - GigE : 32768 bytes
    - Linux : 62.30 MBps
    - EROS Monolithic : 60.80 MBps
    - EROS Factored : 54.42 MBps
Invocation path
Ring Buffer

A Single buffer
- Status
  - Payload Size
  - Next Buffer
- Payload

Buffer Linkages
(Incoming data spans these buffers)
Shared memory
Shared memory region
EROS Monolithic