BROCADE

Performance Challenges In Software Networking

Stephen Hemminger

@networkplumber

Who am I?

- Principal Architect
  Brocade vRouter
  (Vyatta)

- Fellow
  Linux Foundation

- Sequent
  Unix SMP networking

- DPDK
  - #3 contributor
- Linux
  - 10+ year contributor
  - Maintainer
    - Bridge
    - iproute

Agenda

- Myths
- Requirements
- Benchmarks
- Reality

Myths

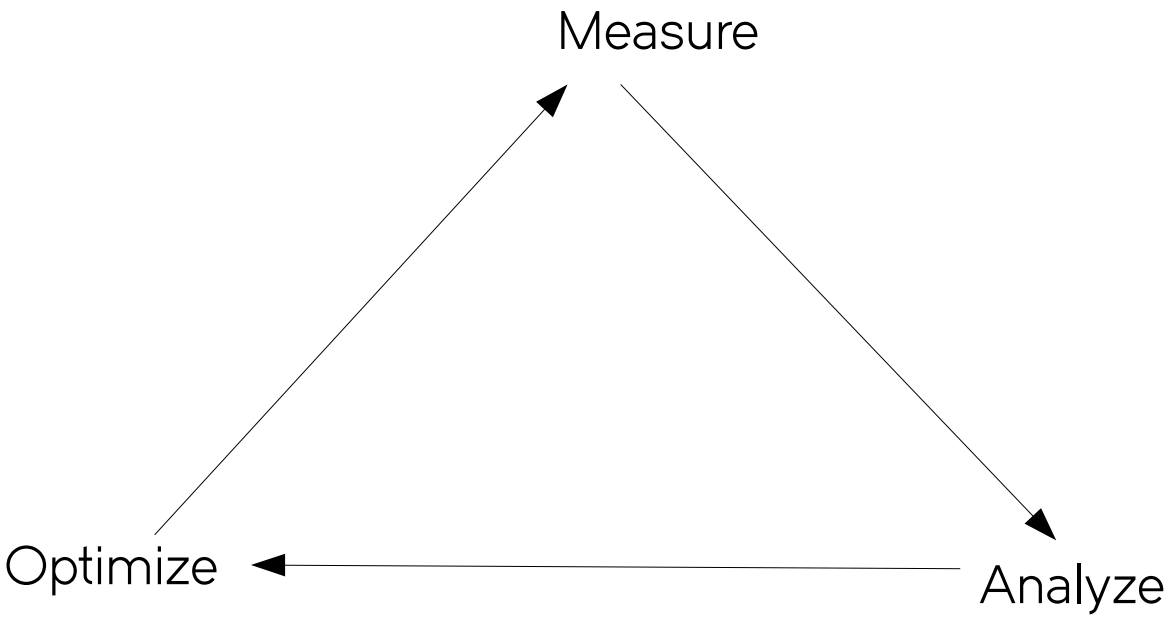
- Software networking can never do:
  - 1Gbit
    - 2008 – Linux, FreeBSD, ...
  - 10Gbit
    - 2013 – DPDK, Netmap, ...
  - 40Gbit
    - 2015 – DPDK, ...
  - 100Gbit
    - 2016?

Hardware vs Software

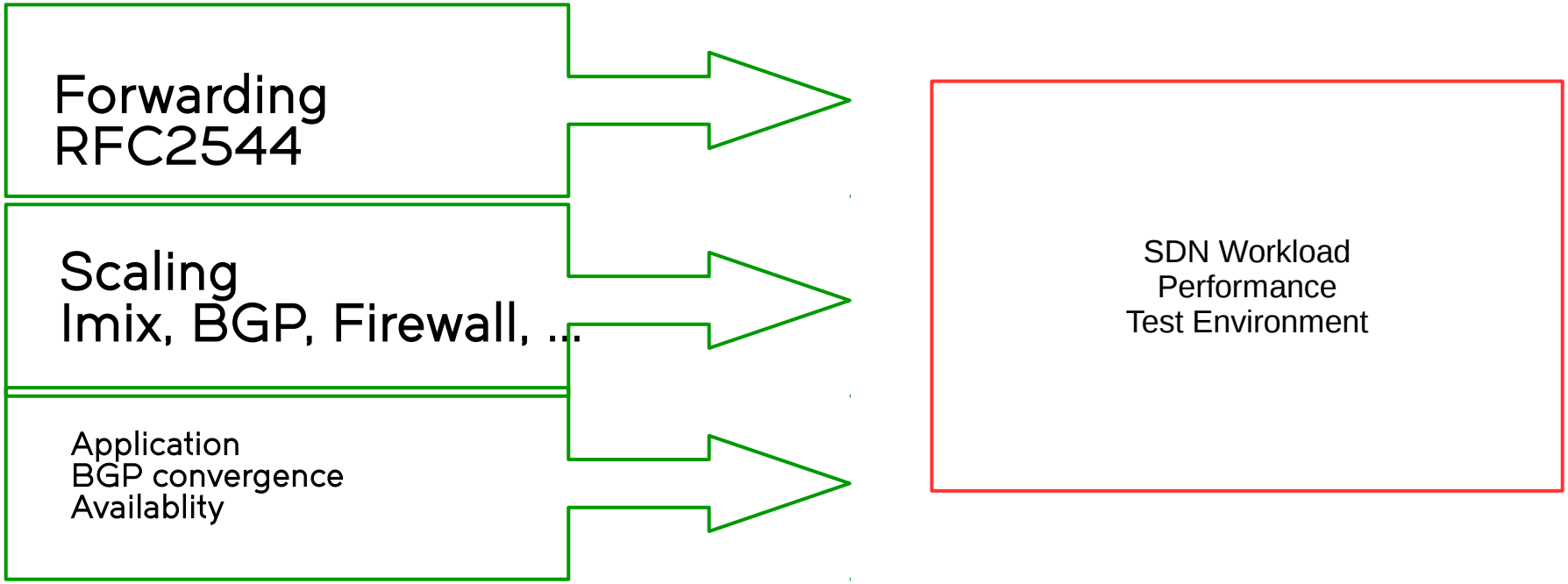
- Clock rate
- TCAM size
- TCAM miss
- Bus transactions

- Clock rate
- Cache size
- Cache misses per packet
- PCI bus operations

Optimization cycle



SDN Measurement



Benchmark vs Reality

- Benchmark
  - random flows
  - 10 or less rules
  - 128GB memory
  - 32 or more CPU

- Reality
  - Burstyflows
  - 1000's of rules
  - 2GB VM
  - 2-4 CPU

System effects

- Data/Control resource sharing
  - CPU cache
  - Background noise
- Power consumption
- Memory footprint
- Virtualization overhead
- Platform differences

Basics

	Sandy Bridge Ivy Bridge	Haswell	Skylake
(bytes/cycle)	4	4	4
L1 Peak Bandwidth	2x16	2x32 load 1x32 store	2x32 load 1x32 store
L2 data access (cycles)	12	11	12
L2 peak Bandwidth	1x32	64	64
Shared L3 access (cycles)	26-31	34	44
L3 peak bandwidth	32	-	32
Data hit in L2 cache	43 – clean hit 60 – modified		

memory is ~70+ ns away (i.e. 2.0 GHz = 140+ cycles)

Source: Intel® 64 and IA-32 Architectures: Optimization Reference Manual

Time Budget

- 10Gbit 64 byte packet
  - 67.2ns = 201 cycles @ 3Ghz
- Cache
  - L3 = 8 ns
  - L2 = 4.3
- Atomic operations
  - Lock = 8.25 ns
  - Lock/Unlock = 16.1

Magic Exlir?

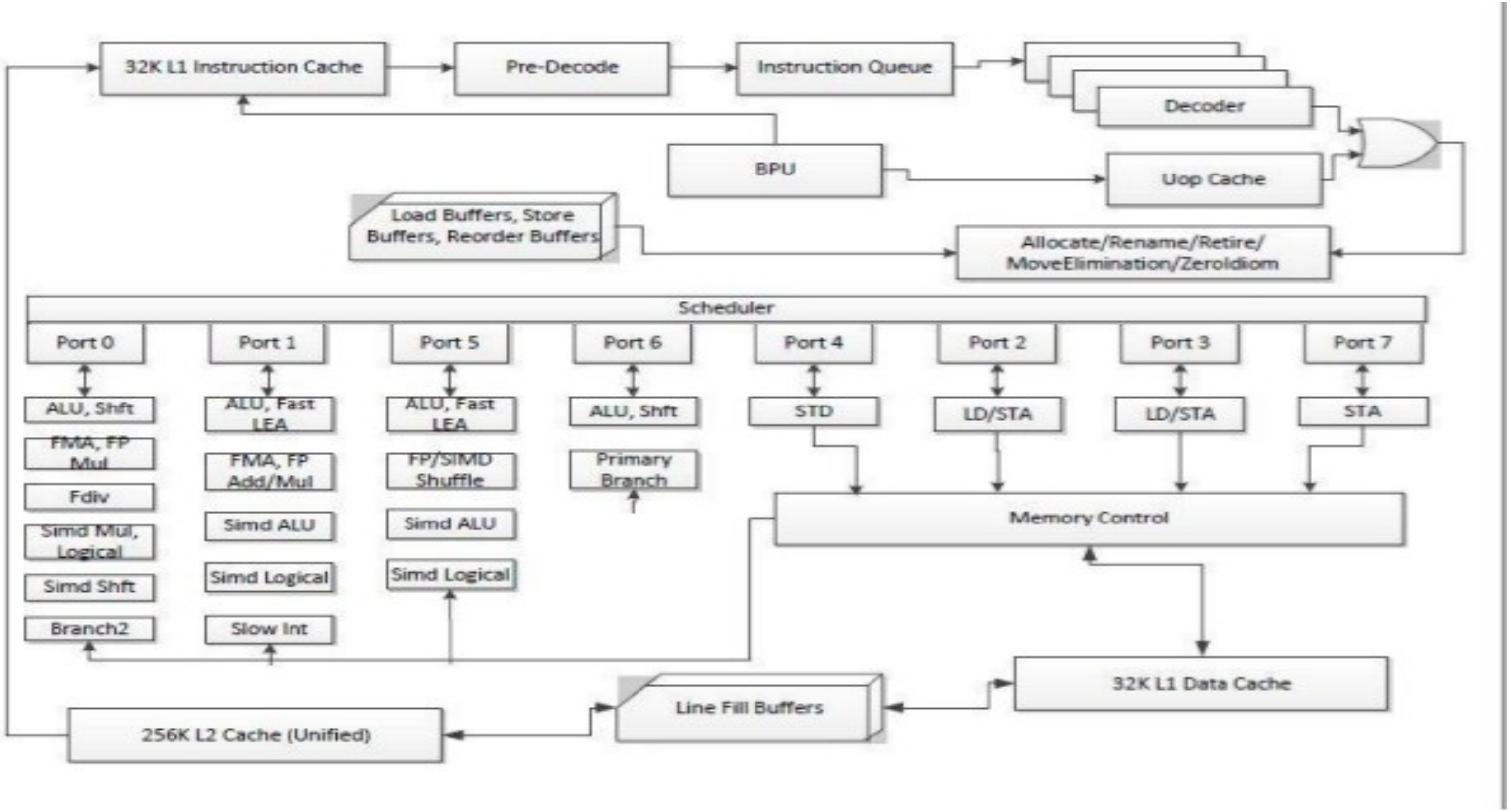
Fast vs Slow

- New software
  - Lockless
  - Single function
  - Tight layering
  - Cache aware
- Legacy software
  - Interrupts
  - Shared resources
  - System calls
  - VM exit

Performance Tradeoffs

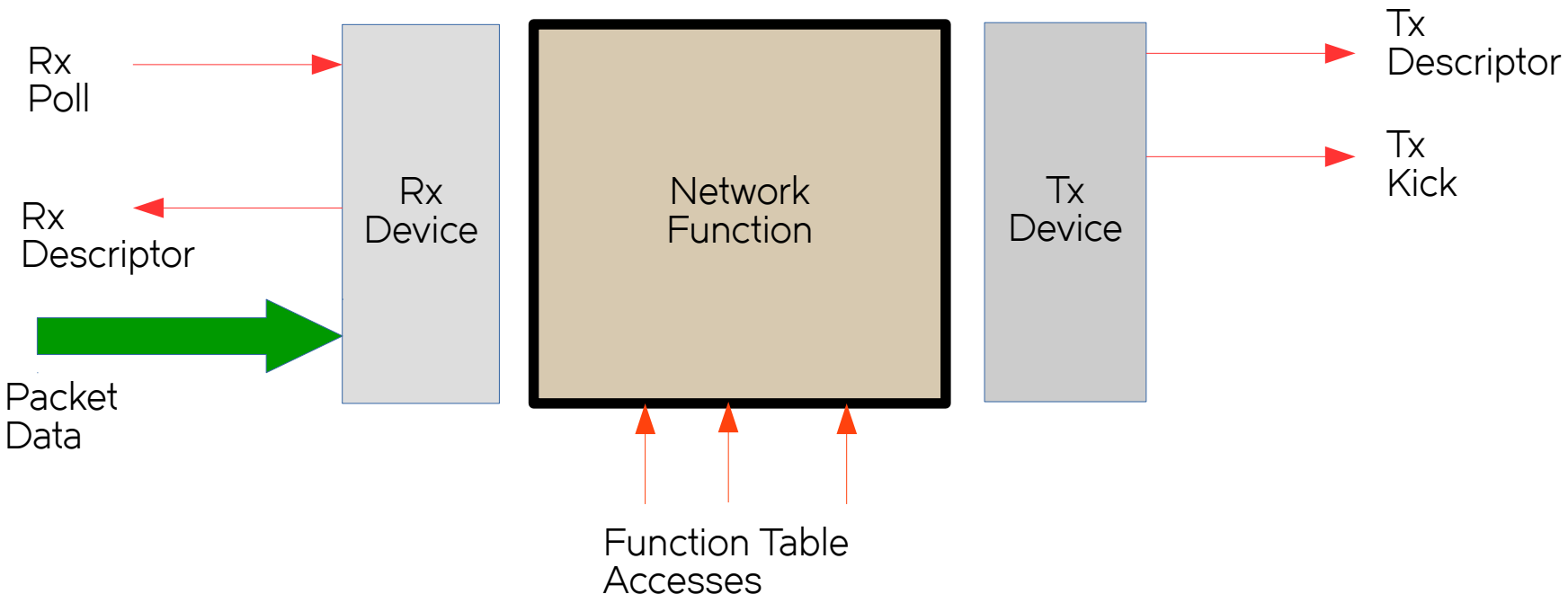
- Bulk operations → Latency
- Lock-less Algorithms → Update speed
  Consistency
- Tight integration → Inflexible
- Polling → CPU utilization
  Power management
- Caching → Memory utilization
  Update overhead

CPU pipeline



Intel Haswell CPU pipeline

Cache flow



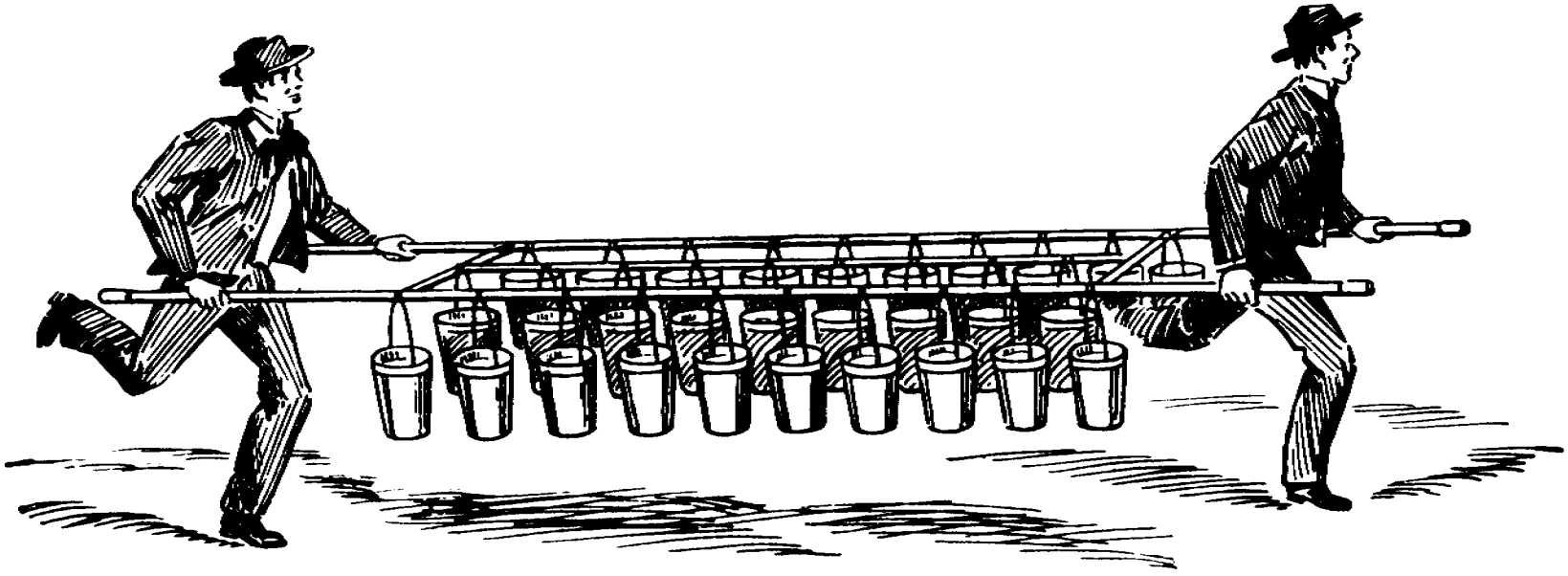
Worst case 7+ cache miss per packet!

Cache Ping/Pong

- Cache line shared between cores
  - Statistics
  - Session state

NFV bucket brigade

Packet batching



New developments

- DPDK
  - Multi-architecture
  - NIC support
  - Packet pipeline
  - ACL
  - LPM
  - ...
- Linux
  - Batched Tx
  - Lockless queue disciplines
  - Memory allocator performance

Conclusions

- Software networking is function of:
  - Algorithims
  - Low level CPU utilization
  - Cache behavior

Questions?

Thank you

Stephen Hemminger
stephen@networkplumber.org
@networkplumber

Next Generation Software Networking

- Openvswitch + DPDK
- Brocade – vRouter
- 6Wind
- FD.io – VPP
- Juniper - Opencontrail
- Huawei - Fusionsphere

BROCADE

Performance Challenges In Software Networking

Stephen Hemminger

@networkplumber

Who am I?

- Principal Architect
  Brocade vRouter
  (Vyatta)

- Fellow
  Linux Foundation

- Sequent
  Unix SMP networking

- DPDK
  - #3 contributor
- Linux
  - 10+ year contributor
  - Maintainer
    - Bridge
    - iproute

Agenda

- Myths
- Requirements
- Benchmarks
- Reality

Myths

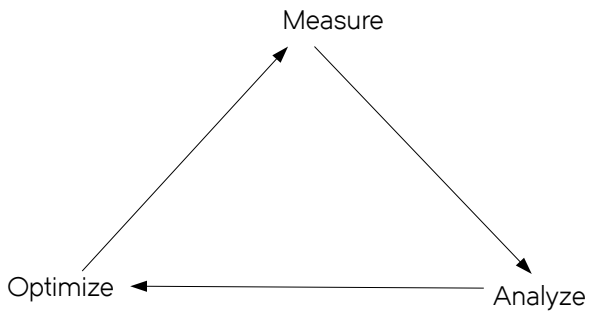
- Software networking can never do:
  - 1Gbit
    - 2008 – Linux, FreeBSD, ...
  - 10Gbit
    - 2013 – DPDK, Netmap, ...
  - 40Gbit
    - 2015 – DPDK, ...
  - 100Gbit
    - 2016?

Hardware vs Software

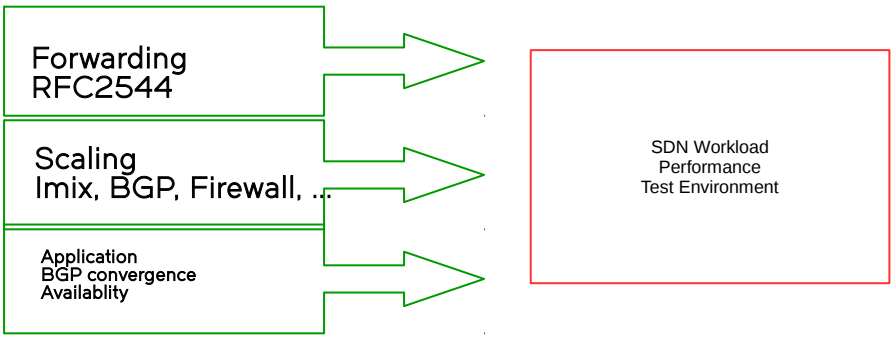
- Clock rate
- TCAM size
- TCAM miss
- Bus transactions

- Clock rate
- Cache size
- Cache misses per packet
- PCI bus operations

Optimization cycle



SDN Measurement



Benchmark vs Reality

- Benchmark
  - random flows
  - 10 or less rules
  - 128GB memory
  - 32 or more CPU

- Reality
  - Burstyflows
  - 1000's of rules
  - 2GB VM
  - 2-4 CPU

System effects

- Data/Control resource sharing
  - CPU cache
  - Background noise
- Power consumption
- Memory footprint
- Virtualization overhead
- Platform differences

Basics

	Sandy Bridge Ivy Bridge	Haswell	Skylake
(bytes/cycle)	4	4	4
L1 Peak Bandwidth	2x16	2x32 load 1x32 store	2x32 load 1x32 store
L2 data access (cycles)	12	11	12
L2 peak Bandwidth	1x32	64	64
Shared L3 access (cycles)	26-31	34	44
L3 peak bandwidth	32	-	32
Data hit in L2 cache	43 – clean hit 60 – modified		

memory is ~70+ ns away (i.e. 2.0 GHz = 140+ cycles)

Source: Intel® 64 and IA-32 Architectures: Optimization Reference Manual

Time Budget

- 10Gbit 64 byte packet
  - 67.2ns = 201 cycles @ 3Ghz
- Cache
  - L3 = 8 ns
  - L2 = 4.3
- Atomic operations
  - Lock = 8.25 ns
  - Lock/Unlock = 16.1

Magic Exlir?

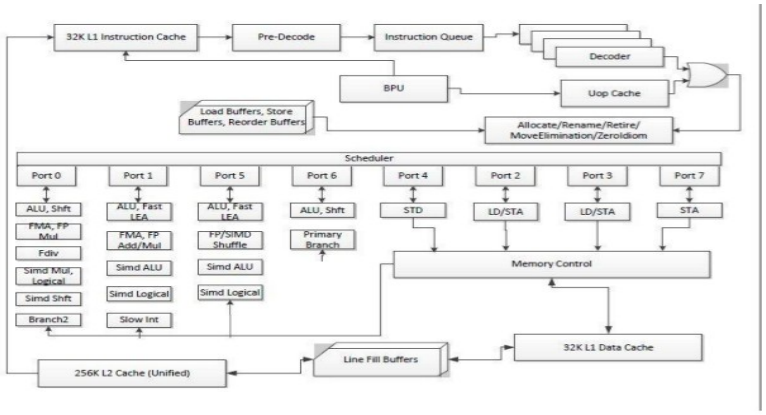
Fast vs Slow

- New software
  - Lockless
  - Single function
  - Tight layering
  - Cache aware
- Legacy software
  - Interrupts
  - Shared resources
  - System calls
  - VM exit

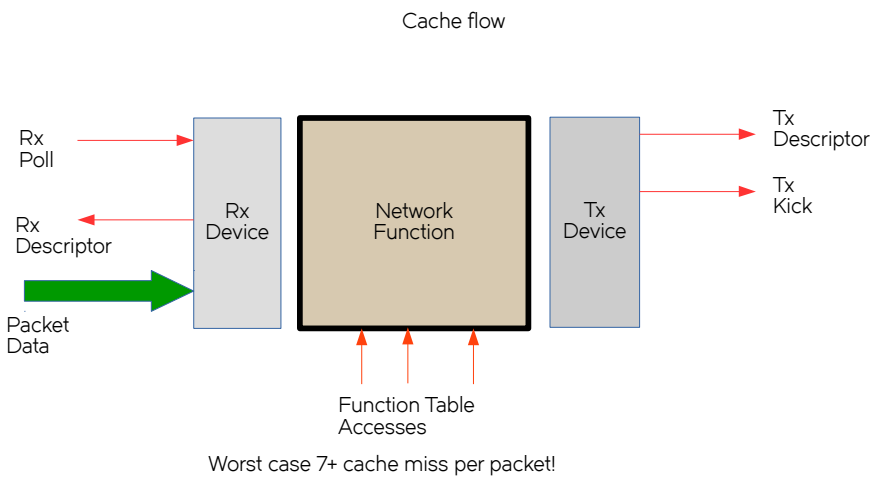
Performance Tradeoffs

- Bulk operations → Latency
- Lock-less Algorithms → Update speed, Consistency
- Tight integration → Inflexible
- Polling → CPU utilization, Power management
- Caching → Memory utilization, Update overhead

CPU pipeline



Intel Haswell CPU pipeline

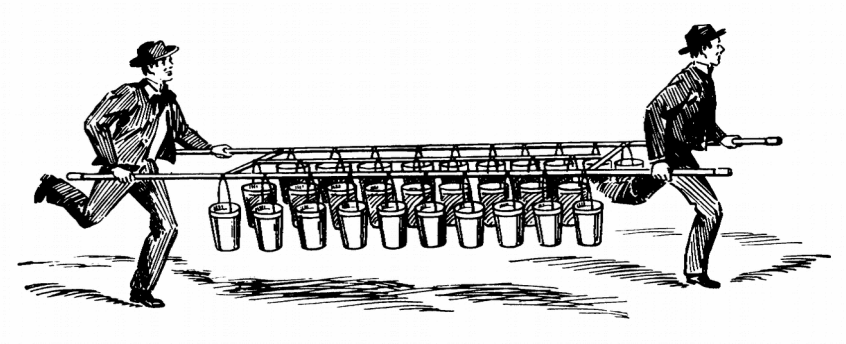

Cache flow
Rx
Poll
Rx
Descriptor
Packet
Data
Rx
Device
Network
Function
Tx
Device
Tx
Descriptor
Tx
Kick
Function Table
Accesses
Worst case 7+ cache miss per packet!

Cache Ping/Pong

- Cache line shared between cores
  - Statistics
  - Session state


NFV bucket brigade
FIRE DEPARTMENT

Packet batching



New developments

- DPDK
  - Multi-architecture
  - NIC support
  - Packet pipeline
  - ACL
  - LPM
  - ...
- Linux
  - Batched Tx
  - Lockless queue disciplines
  - Memory allocator performance

Conclusions

- Software networking is function of:
  - Algorithims
  - Low level CPU utilization
  - Cache behavior

Questions?

Thank you

Stephen Hemminger
stephen@networkplumber.org
@networkplumber

Next Generation Software Networking

- Openvswitch + DPDK
- Brocade – vRouter
- 6Wind
- FD.io – VPP
- Juniper - Opencontrail
- Huawei - Fusionsphere