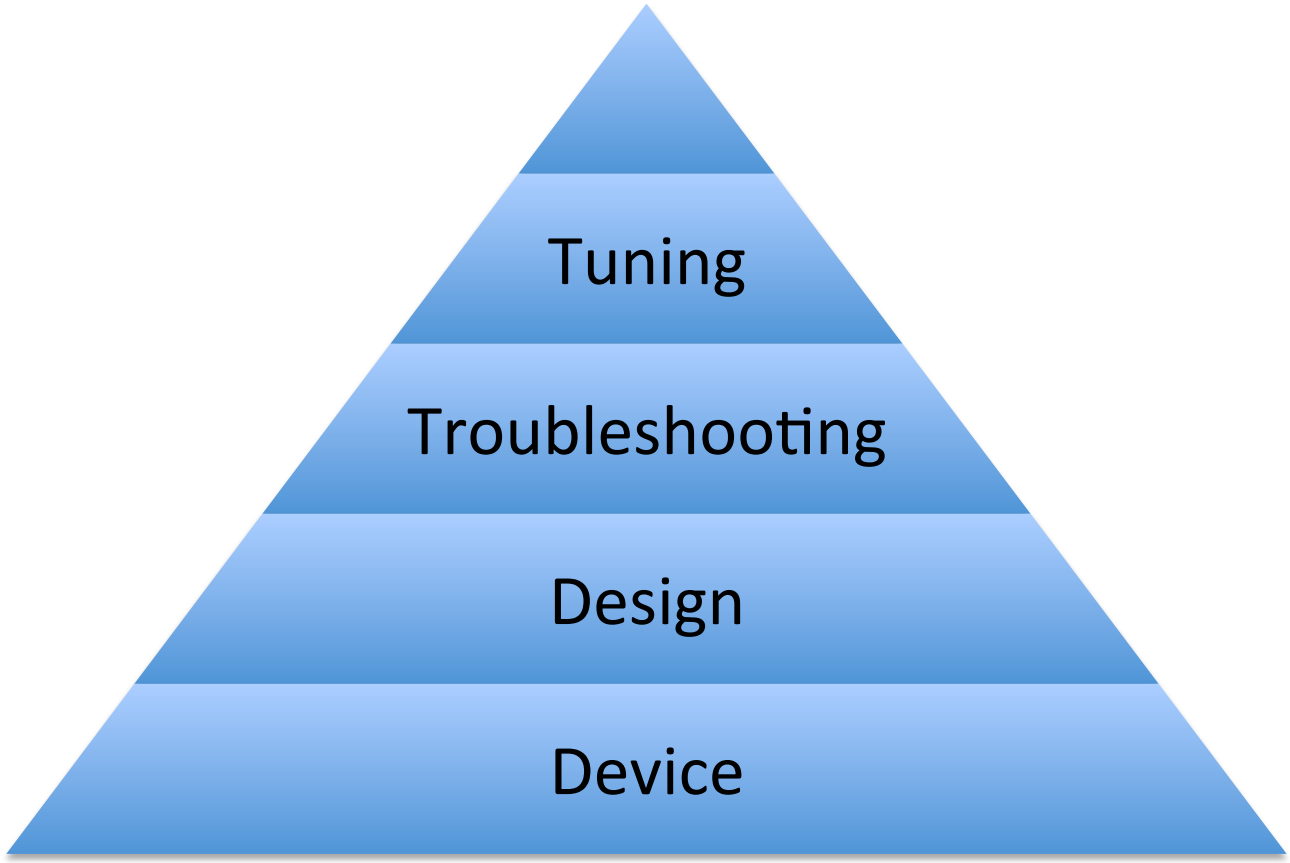
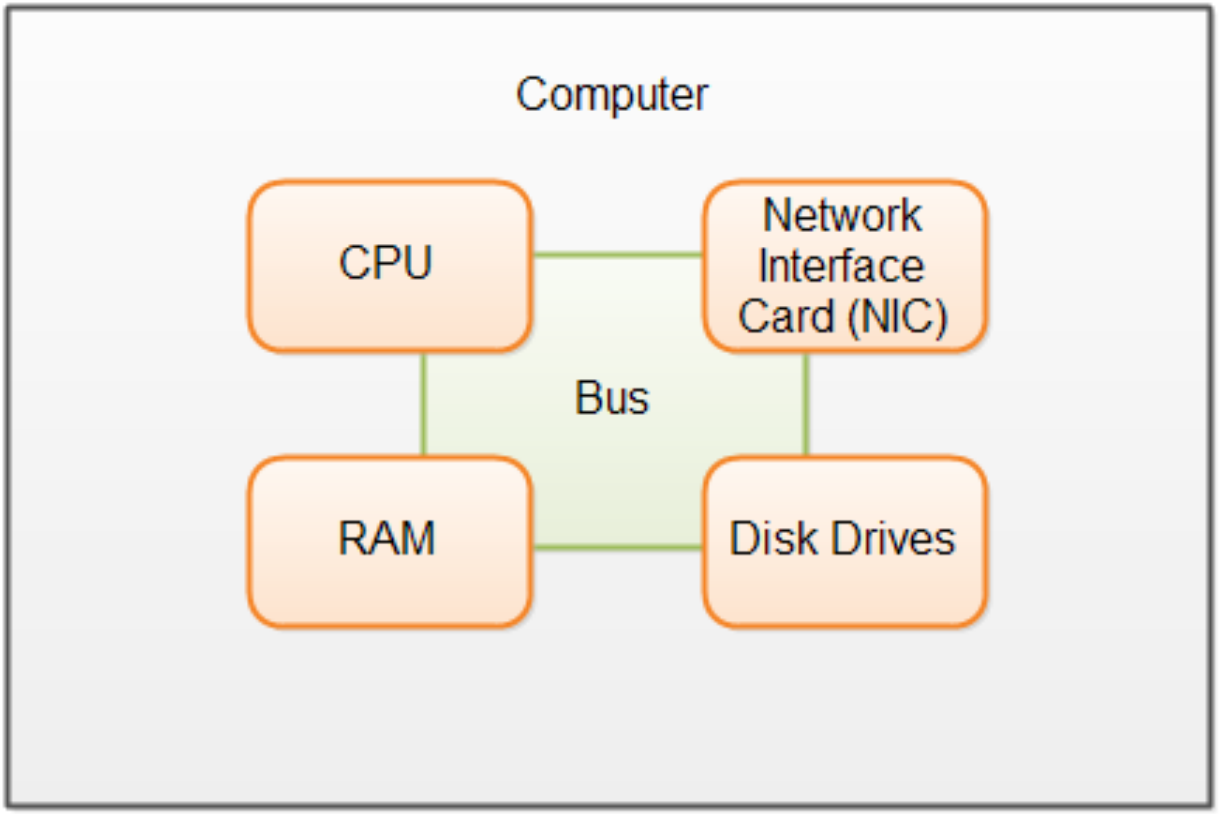
高性能服务器架构设计和调优

千石

Agenda

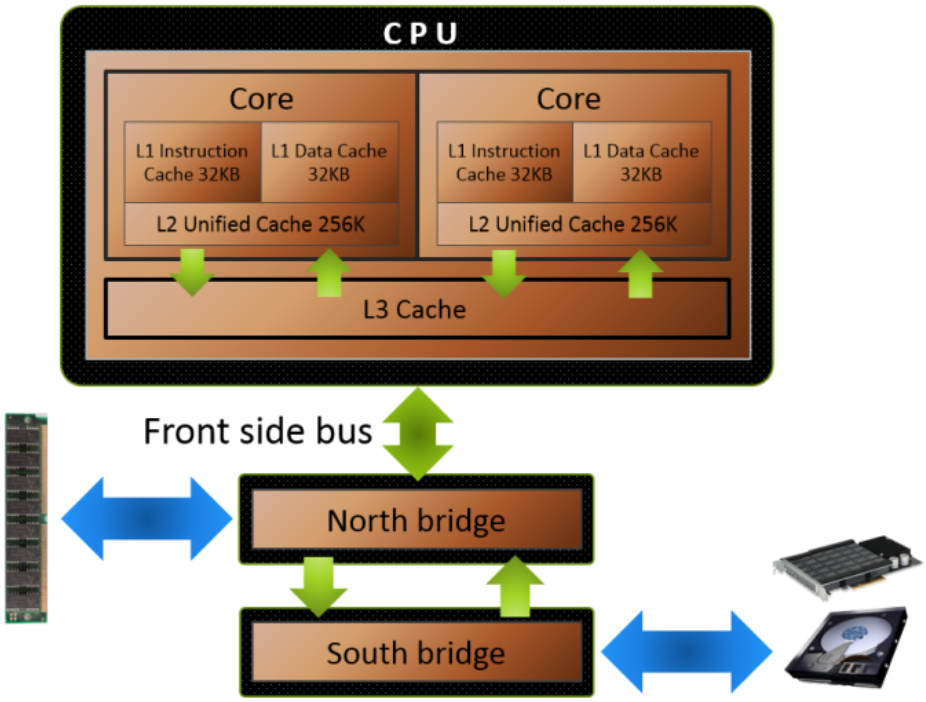


Device

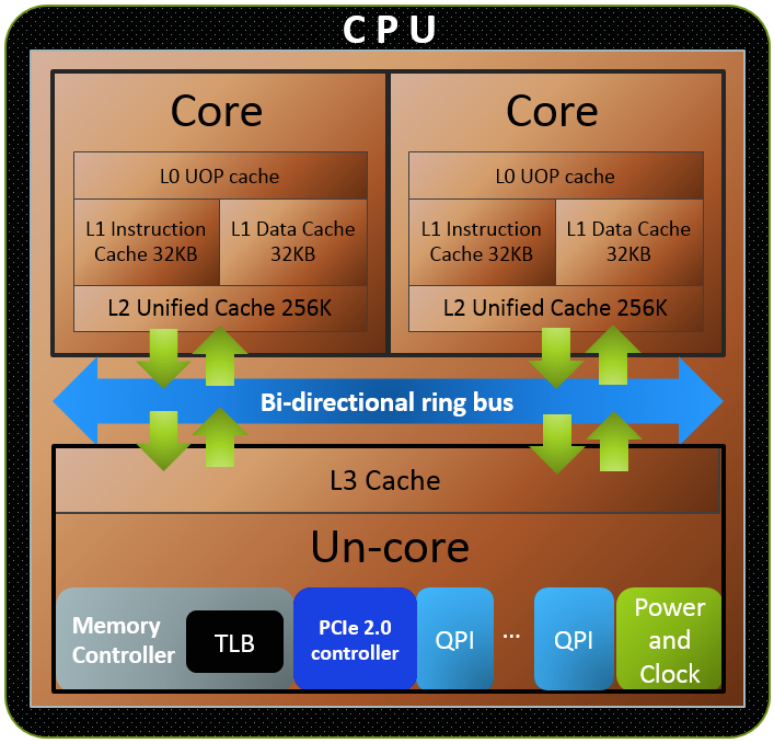

Computer
CPU
Network Interface Card (NIC)
Bus
RAM
Disk Drives

CPU

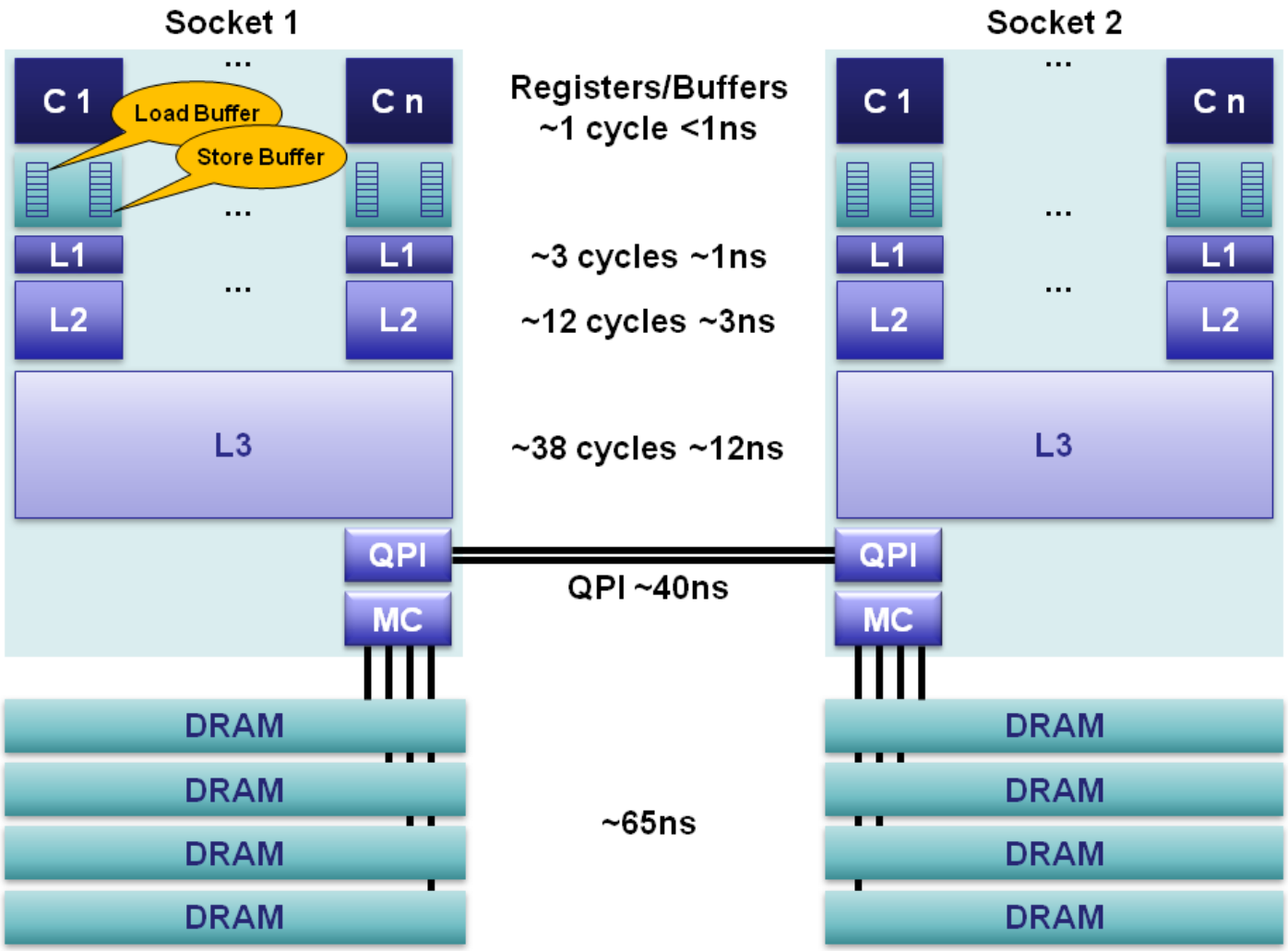
Intel Core -2



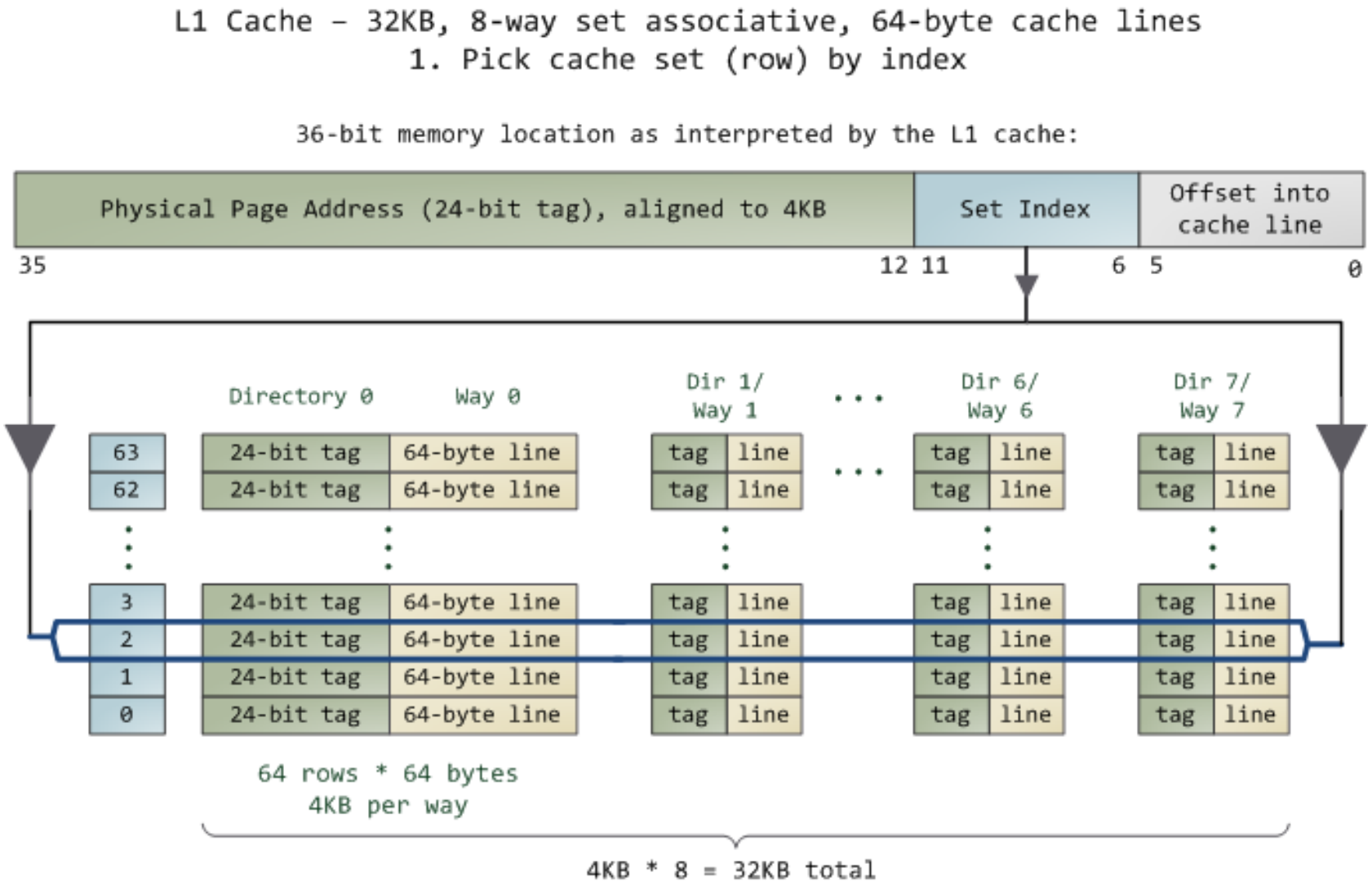
Intel Sandybridge



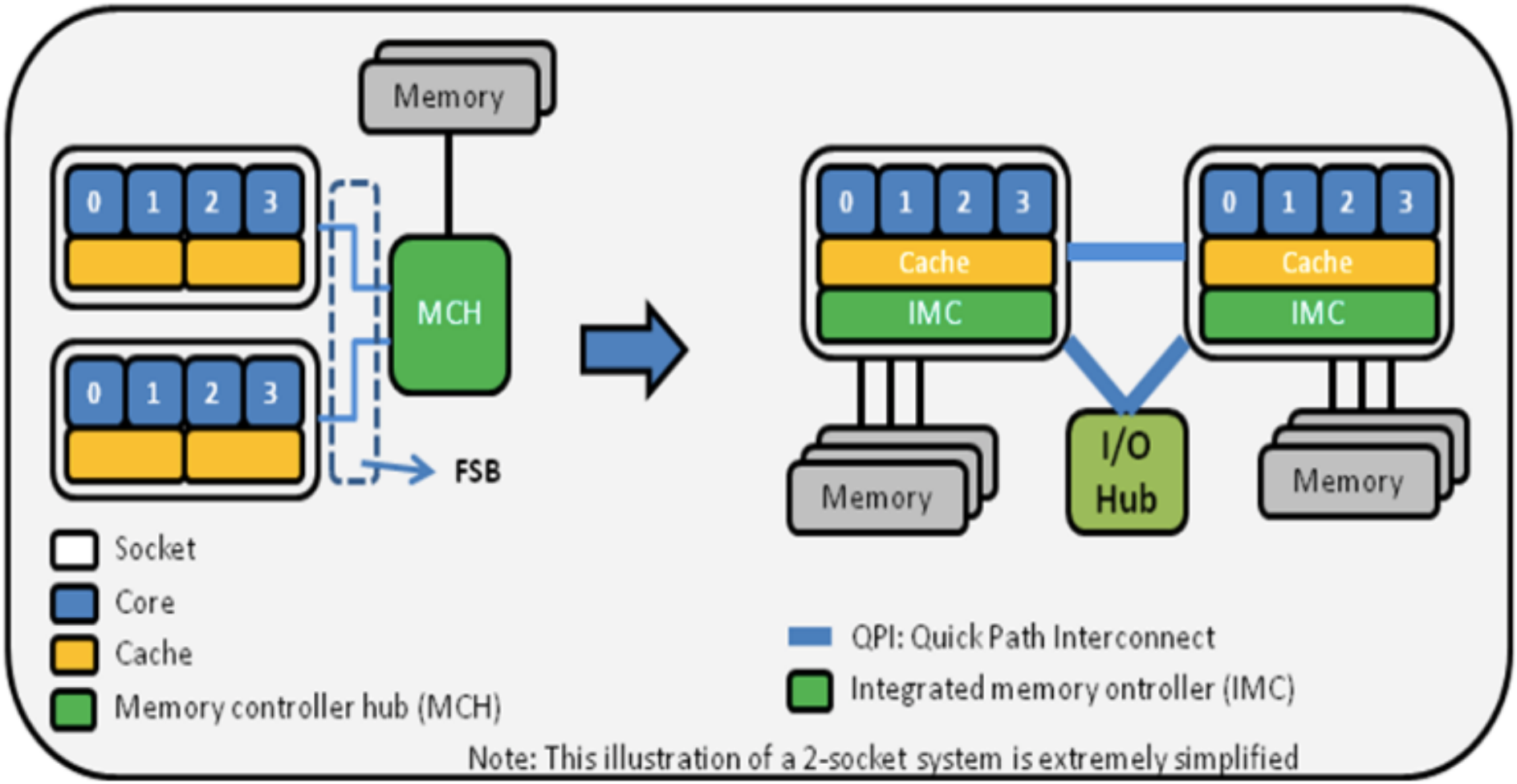
CPU Cache


Socket 1
Socket 2
C 1
C n
Load Buffer
Store Buffer
Registers/Buffers
~1 cycle <1ns
L1
~3 cycles ~1ns
L2
~12 cycles ~3ns
L3
~38 cycles ~12ns
QPI
QPI ~40ns
MC
DRAM
~65ns

CPU Cache line

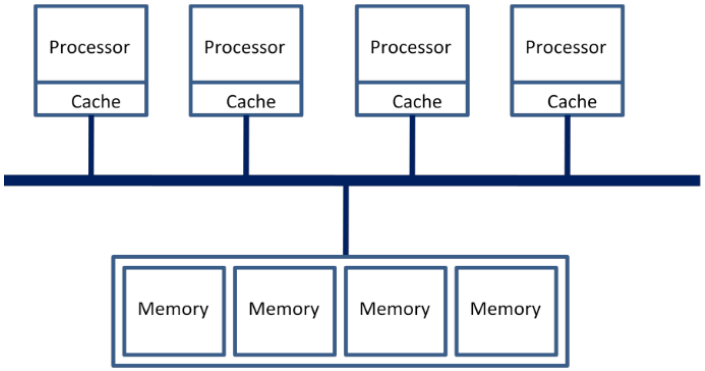


Memory Controler

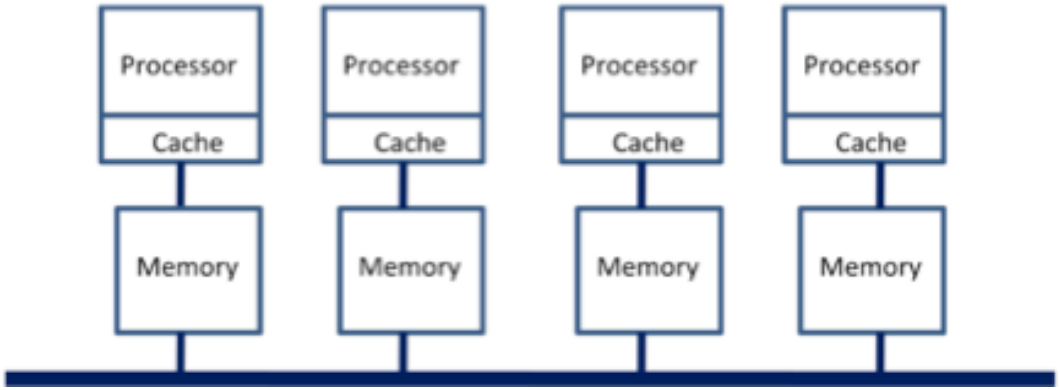


UMA/NUMA

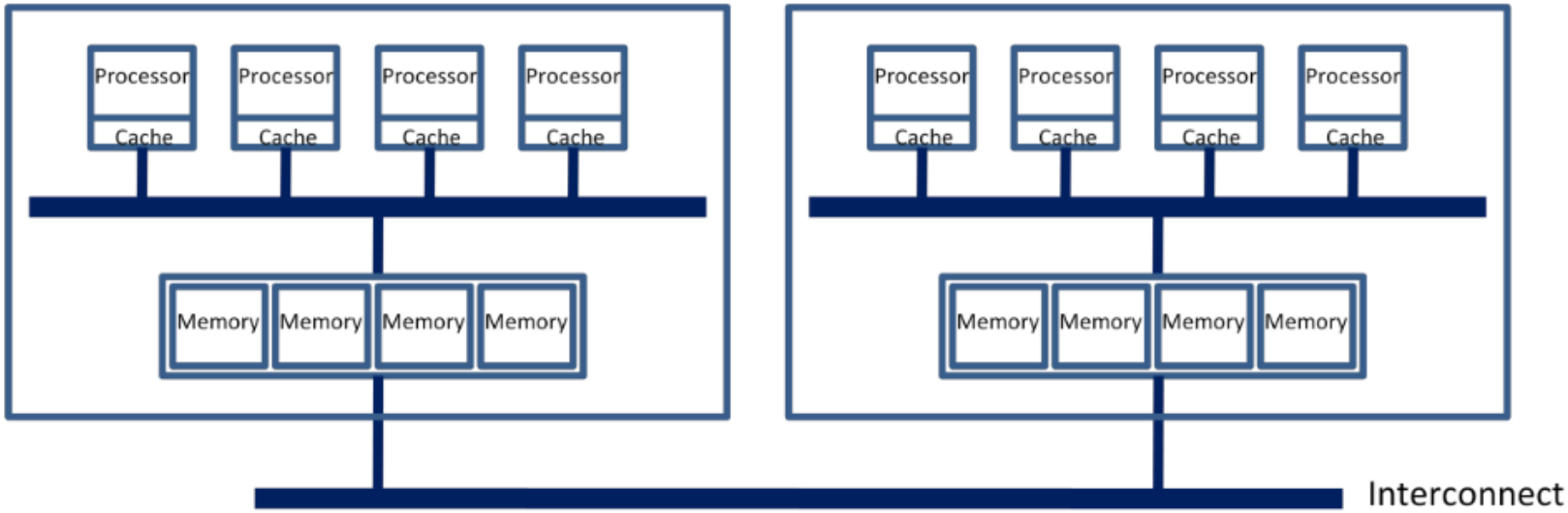
UMA



NUMA

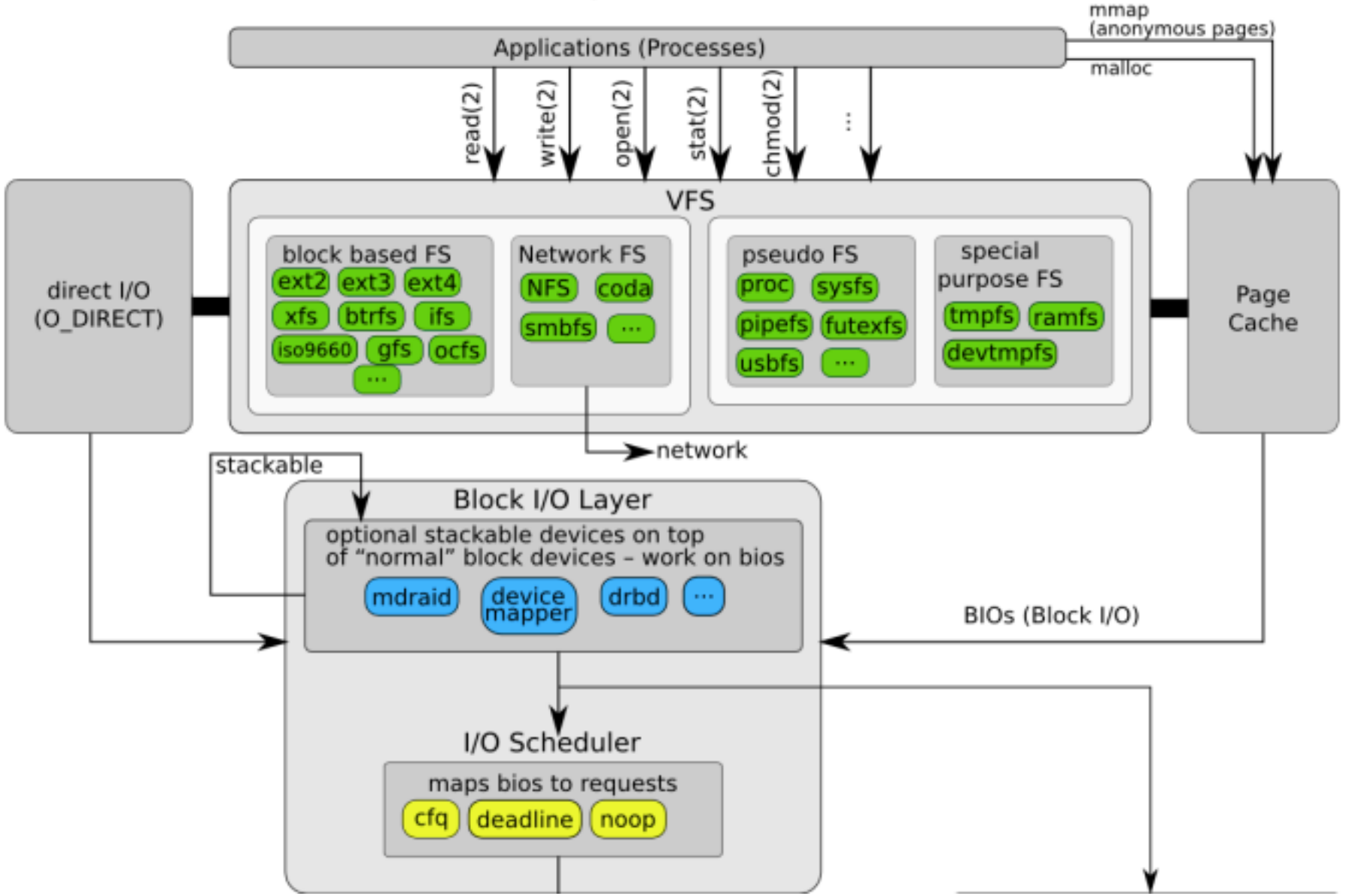


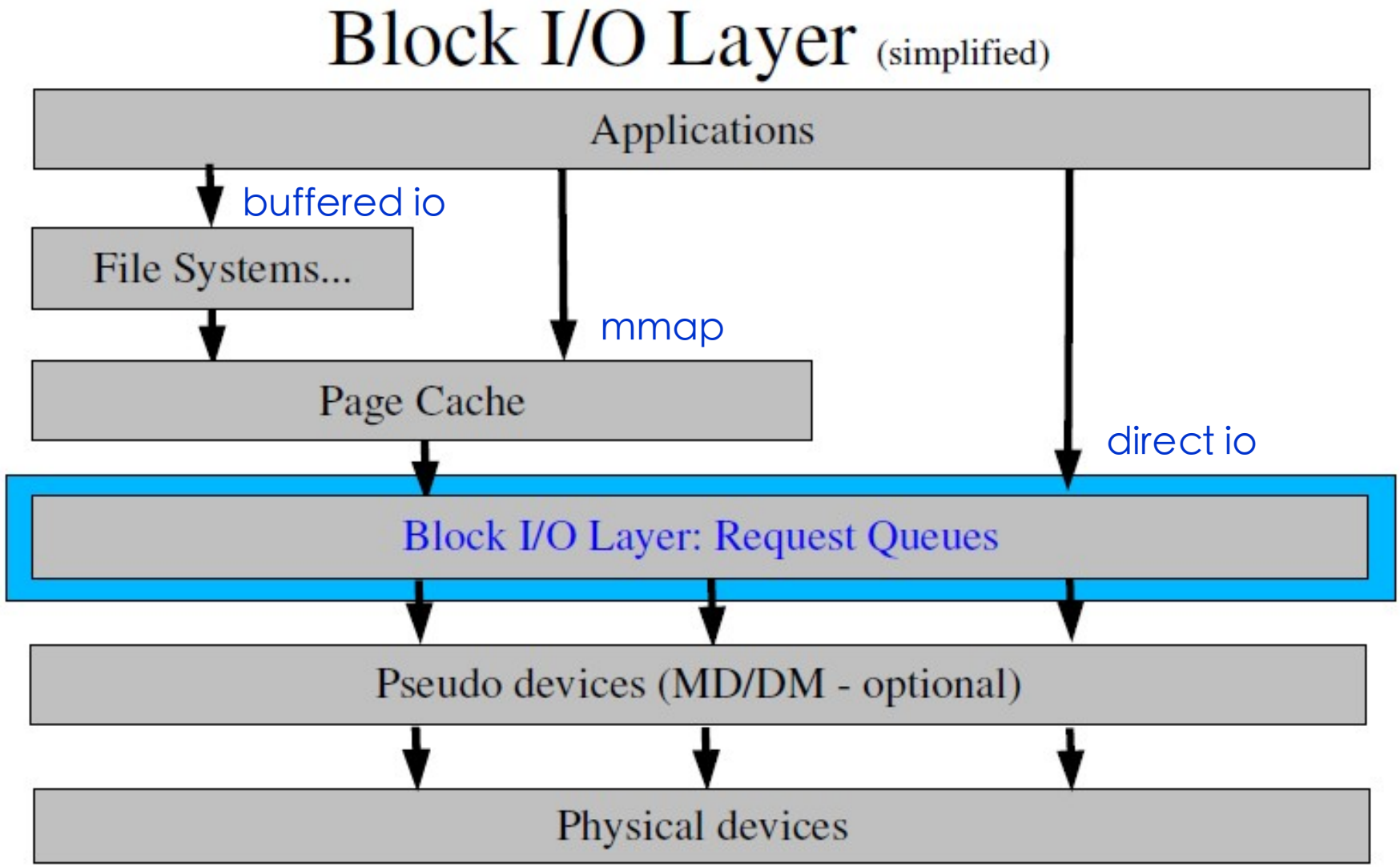
Mix architectures



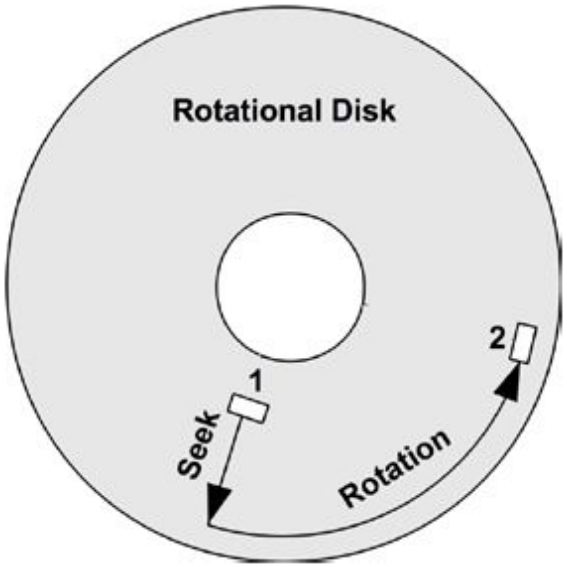
The Linux I/O Stack Diagram

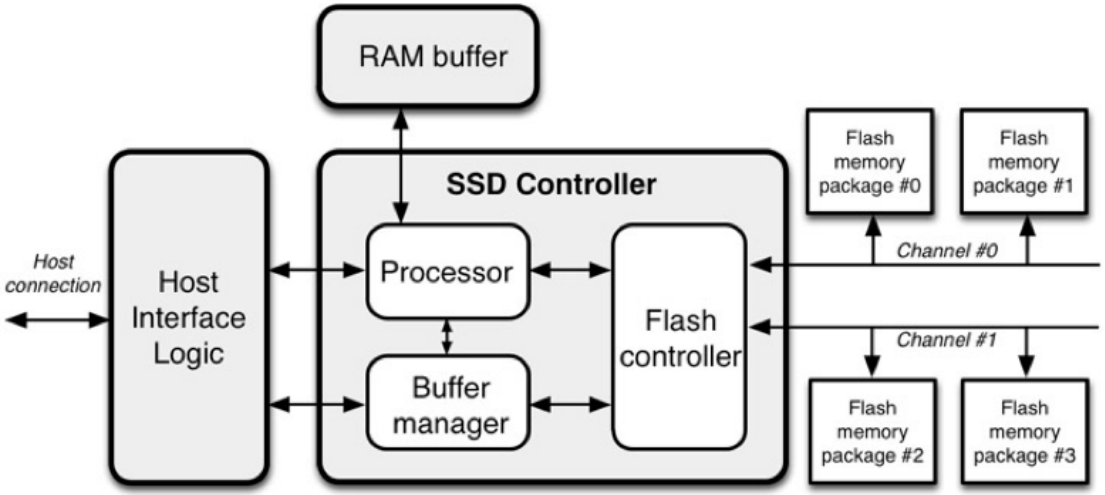
version 1.0, 2012-06-20
outlines the Linux I/O stack as of Kernel version 3.3




Block I/O Layer (simplified)
Applications
buffered io
File Systems...
mmap
Page Cache
direct io
Block I/O Layer: Request Queues
Pseudo devices (MD/DM - optional)
Physical devices

硬盘到SSD


Rotational Disk
1
2
Seek
Rotation


RAM buffer
SSD Controller
Host connection
Host Interface Logic
Processor
Buffer manager
Flash controller
Flash memory package #0
Flash memory package #1
Flash memory package #2
Flash memory package #3
Channel #0
Channel #1

Writing data to a solid-state drive

1. Initial configuration

Block 1000 (data)

PPN	data
0	x
1	y
2	z
3	

Block 2000 (free)

PPN	data
0	
1	
2	
3	

- Initially, block 2000 is free and block 1000 has three used pages at PPN = 0, 1, and 2 (Physical Page Number), and one free page at PPN = 3.

2. Writing a page

Block 1000 (data)

PPN	data
0	x
1	y
2	z
3	x'

Block 2000 (free)

PPN	data
0	
1	
2	
3	

- The data in block 1000 at PPN = 0 gets updated and becomes x'.
- Since pages cannot be overwritten, the page that contains x becomes stale (PPN = 0), and the new version of the data is stored in a free page, at PPN = 3.

3. Erasing a block (garbage collection)

Block 1000 (free)

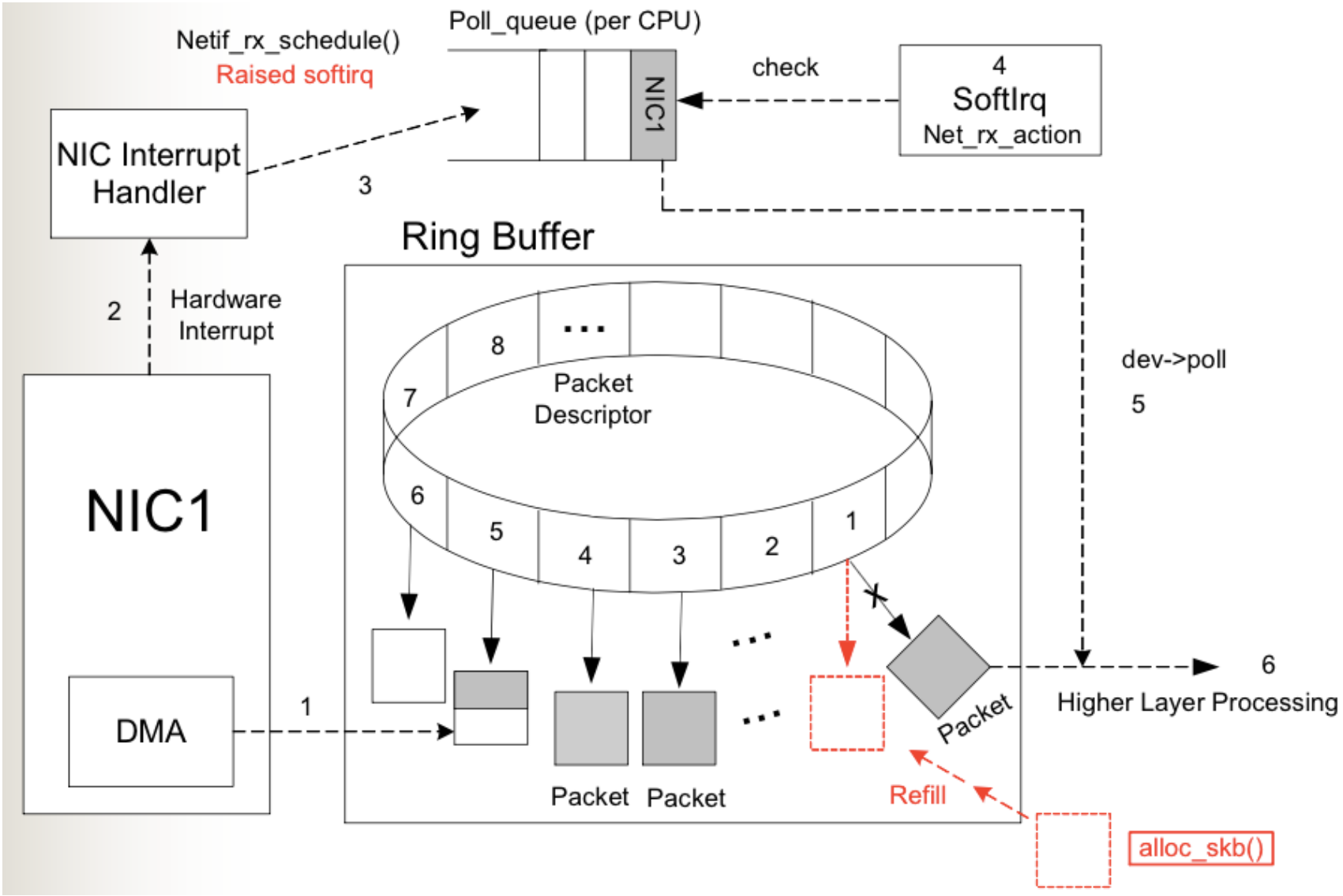
PPN	data
0	
1	
2	
3	

Block 2000 (data)

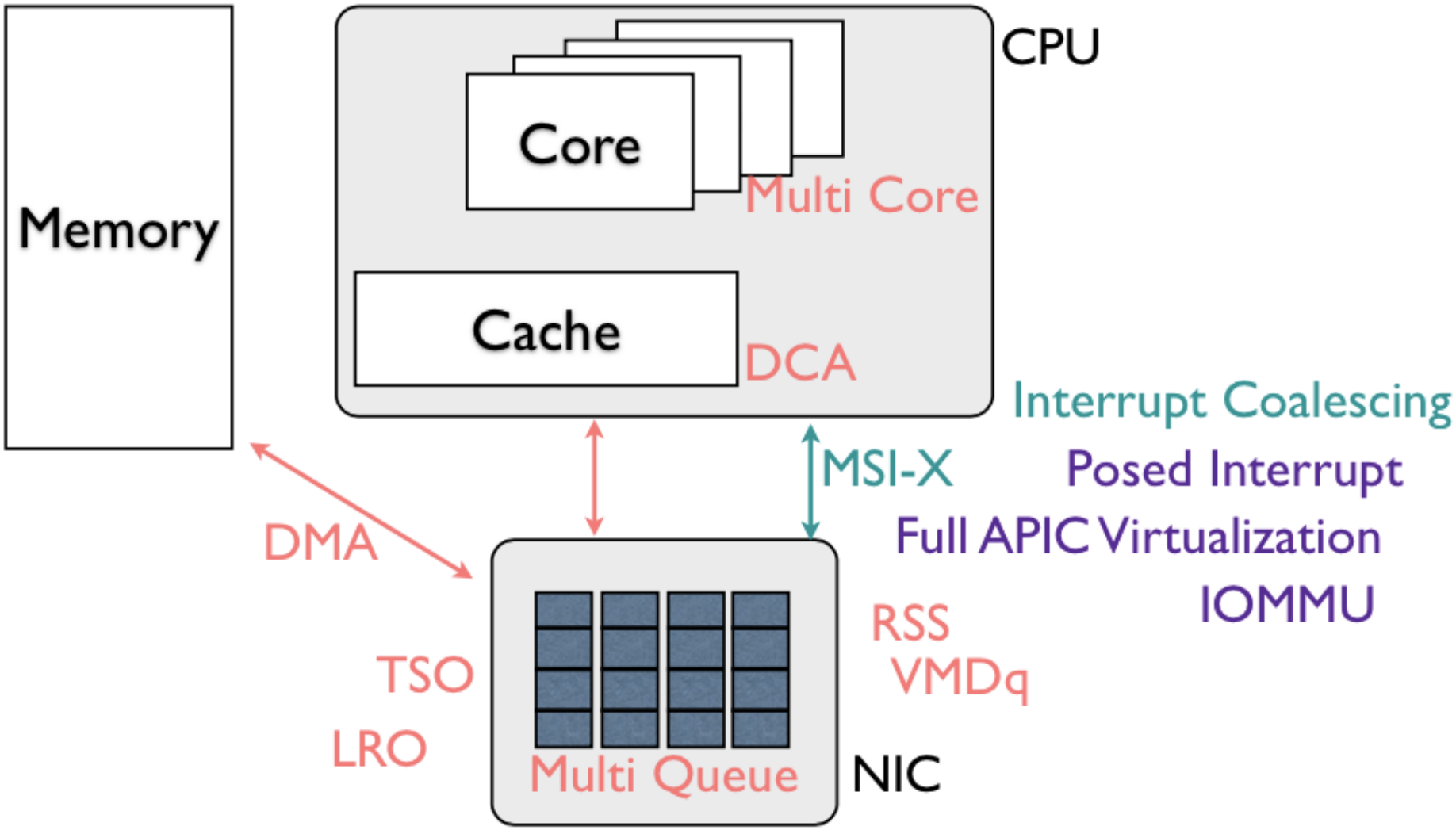
PPN	data
1	y
2	z
3	x'

- The garbage collection process copies all the valid pages from the data block 1000 into the free block 2000, leaving behind the stale pages.
- Block 1000 is erased, which makes it ready to receive new write operations. Blocks can only be erased a limited number of times (P/E cycles) until they wear off and become unusable.

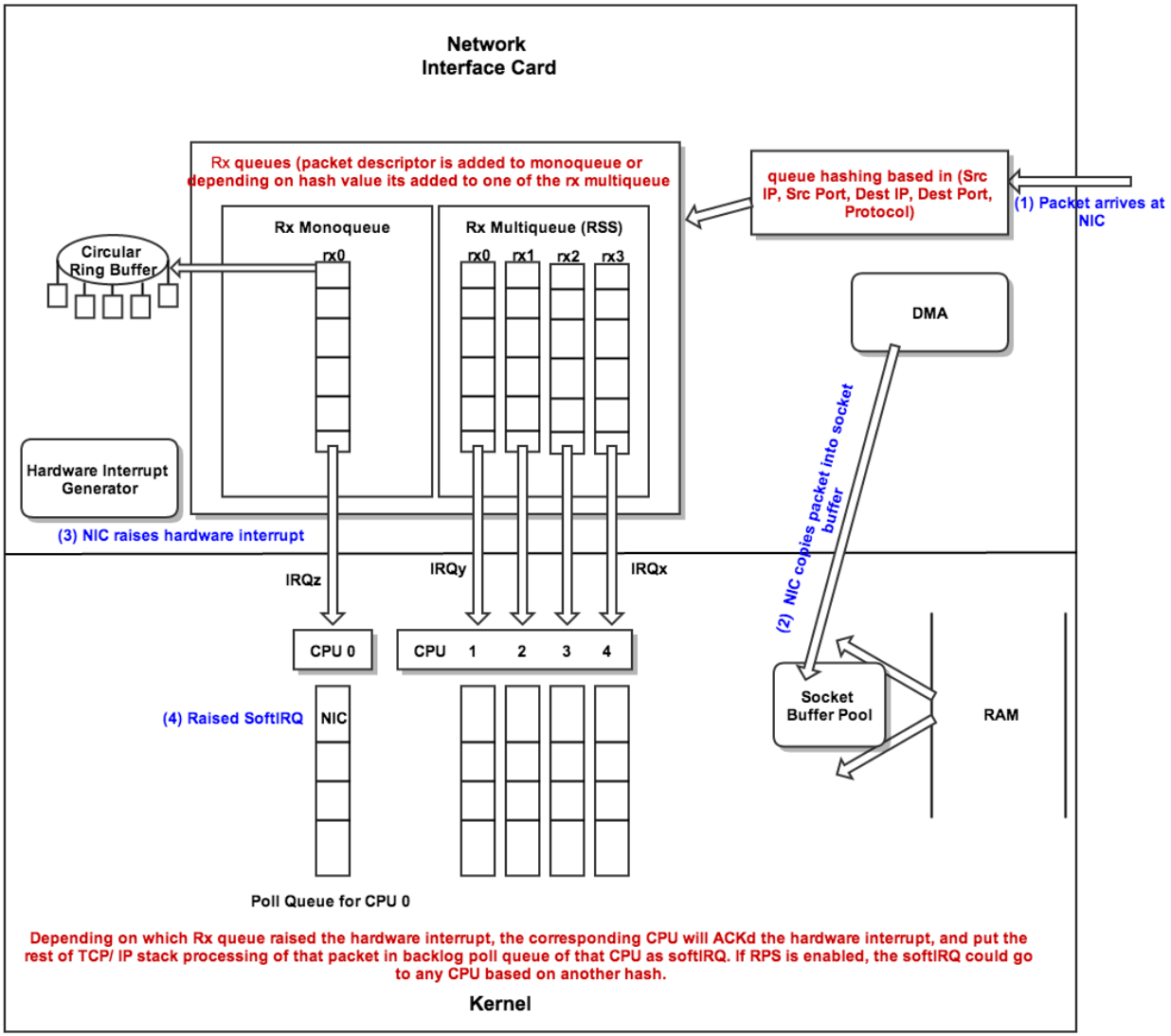
网卡收包


Netif_rx_schedule()
Raised softirq
Poll_queue (per CPU)
NIC1
check
4
SoftIrq
Net_rx_action
NIC Interrupt Handler
3
Ring Buffer
2
Hardware Interrupt
dev->poll
5
NIC1
8
7
6
5
4
3
2
1
Packet Descriptor
DMA
1
Packet
Packet
Packet
6
Higher Layer Processing
Refill
alloc_skb()

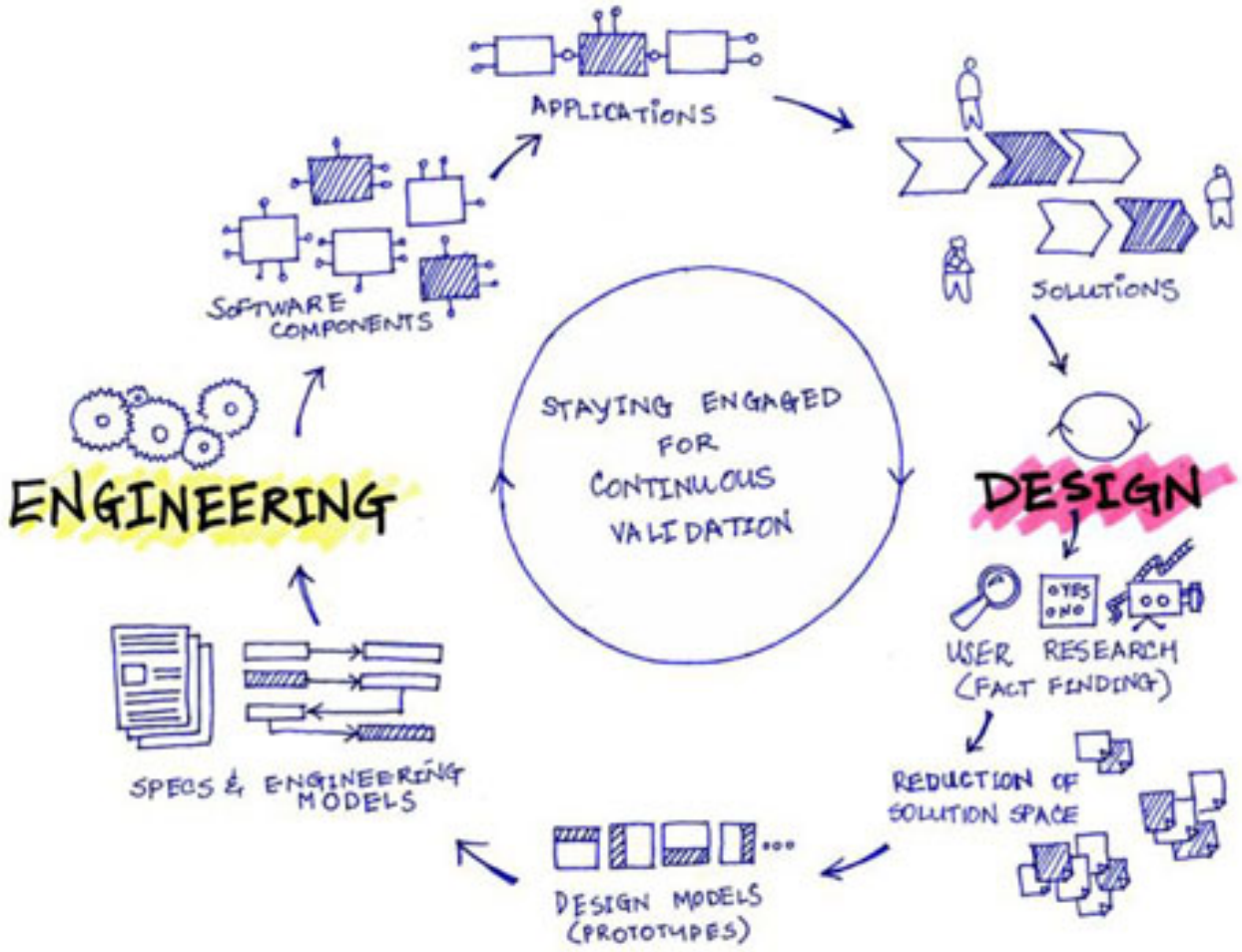
网卡多队列


Memory
CPU
Core
Multi Core
Cache
DCA
Interrupt Coalescing
MSI-X
Posed Interrupt
DMA
Full APIC Virtualization
IOMMU
RSS
VMDq
TSO
LRO
Multi Queue
NIC

网卡多队列绑定CPU


Network
Interface Card
Rx queues (packet descriptor is added to monoqueue or depending on hash value its added to one of the rx multiqueue
queue hashing based in (Src IP, Src Port, Dest IP, Dest Port, Protocol)
(1) Packet arrives at NIC
Rx Monoqueue
rx0
Rx Multiqueue (RSS)
rx0 rx1 rx2 rx3
Circular Ring Buffer
DMA
Hardware Interrupt Generator
(3) NIC raises hardware interrupt
(2) NIC copies packet into socket buffer
IRQz
IRQy
IRQx
CPU 0
CPU 1 2 3 4
(4) Raised SoftIRQ
NIC
Socket Buffer Pool
RAM
Poll Queue for CPU 0
Depending on which Rx queue raised the hardware interrupt, the corresponding CPU will ACKd the hardware interrupt, and put the rest of TCP/ IP stack processing of that packet in backlog poll queue of that CPU as softIRQ. If RPS is enabled, the softIRQ could go to any CPU based on another hash.
Kernel

Design

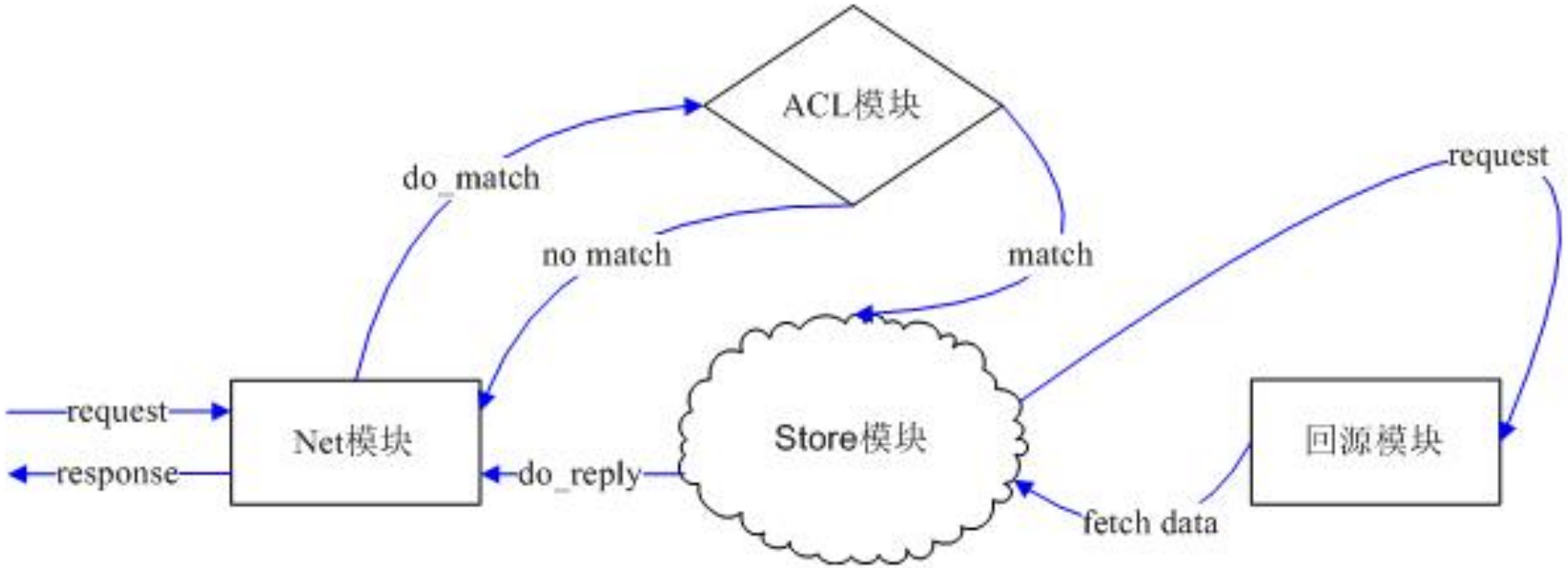

APPLICATIONS
SOLUTIONS
SOFTWARE COMPONENTS
STAYING ENGAGED FOR CONTINUOUS VALIDATION
ENGINEERING
DESIGN
USER RESEARCH (FACT FINDING)
SPECS & ENGINEERING MODELS
REDUCTION OF SOLUTION SPACE
DESIGN MODELS (PROTOTYPES)

软件性能设计原则：不能成为硬件瓶颈

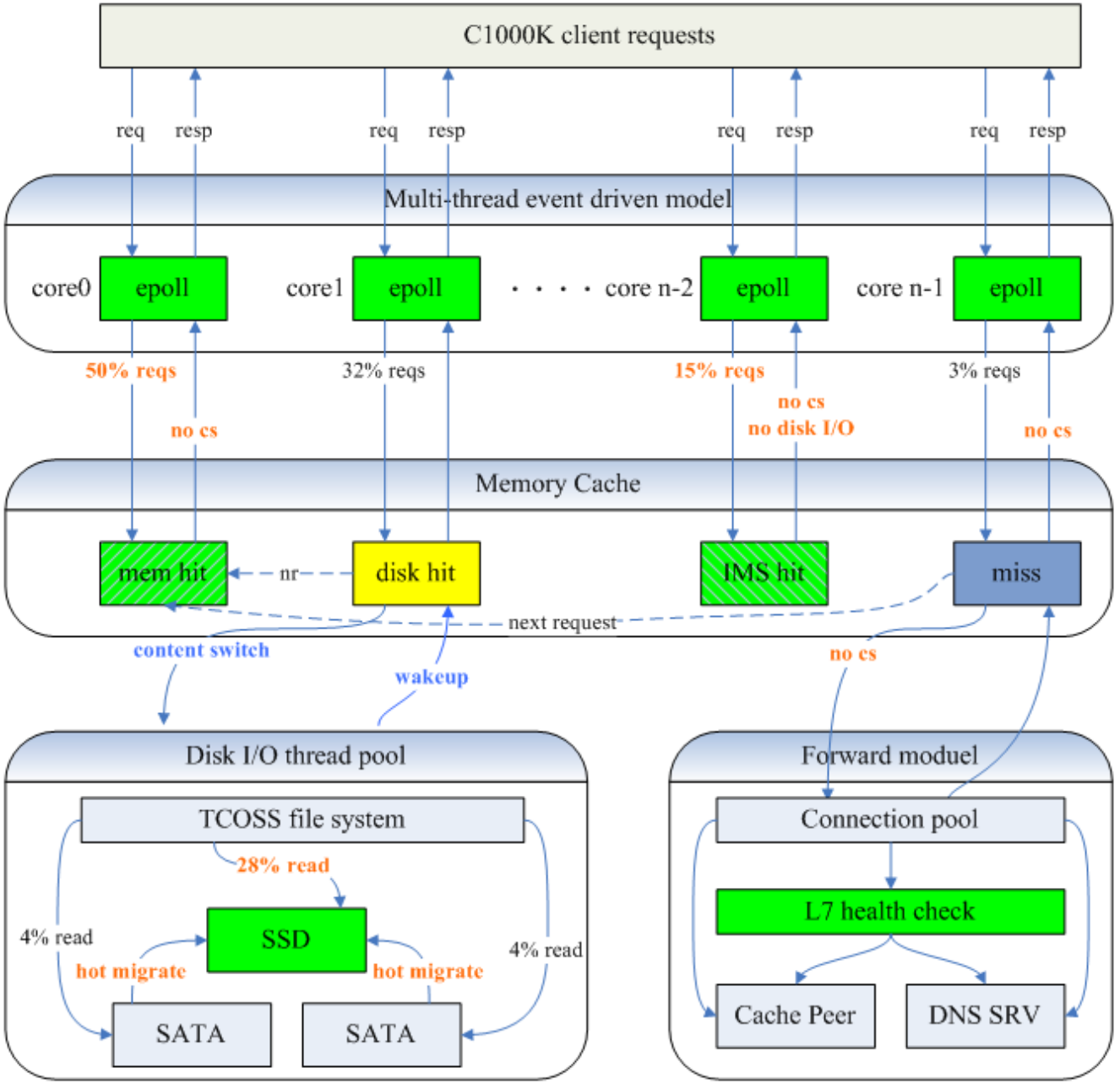
- 均衡使用CPU多核处理能力
- 高效合理地使用和控制内存
- 最大化磁盘IOPS和吞吐，异步化处理
- 小包跑满万兆网卡，中断平衡

CDN Cache系统模型

- Net模块：　支持大并发，跑满万兆网卡
- ACL模块：　高效匹配，减少CPU消耗
- Store模块：　提高命中率，高效利用磁盘IOPS
- 回源模块：　L7-check，长连接保持



Swift -- High Performance Web Cache Arichitecture

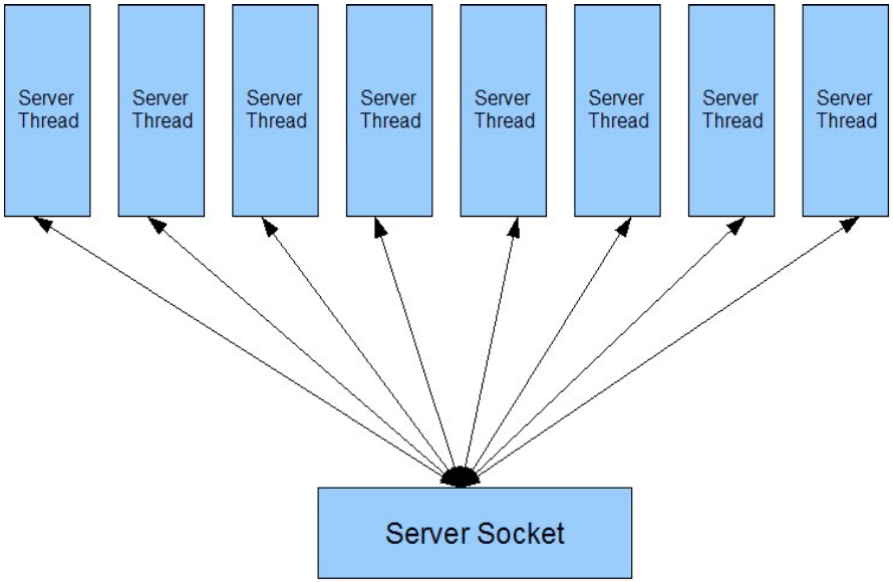


Net 模块

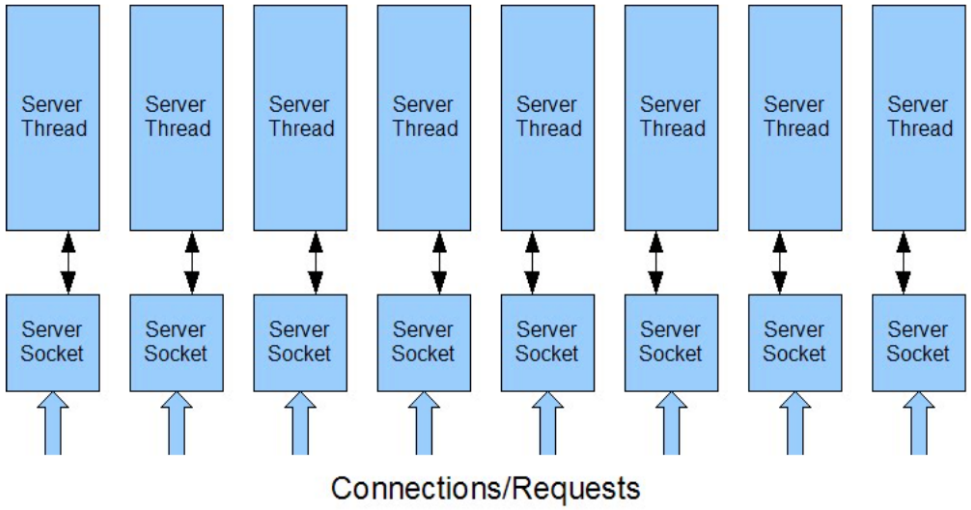
- I/O模型
  - epoll + O_NONBLOCK
- TCP选项
  - TCP_DEFER_ACCEPT / TCP_SYNCNT
  - TCP_CORK / TCP_NODELAY / TCP_QUICKACK
- 收发包的方式
  - RSS, SMP_AFFNITY
  - SO_REUSEPORT

SO_REUSEPORT

- listen单个fd存在accept竞争问题

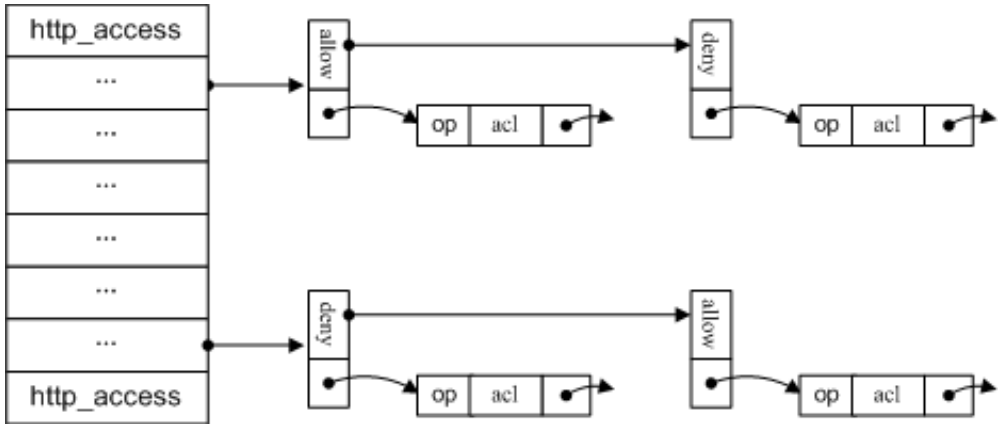


- SO_REUSEPORT listen多个fd

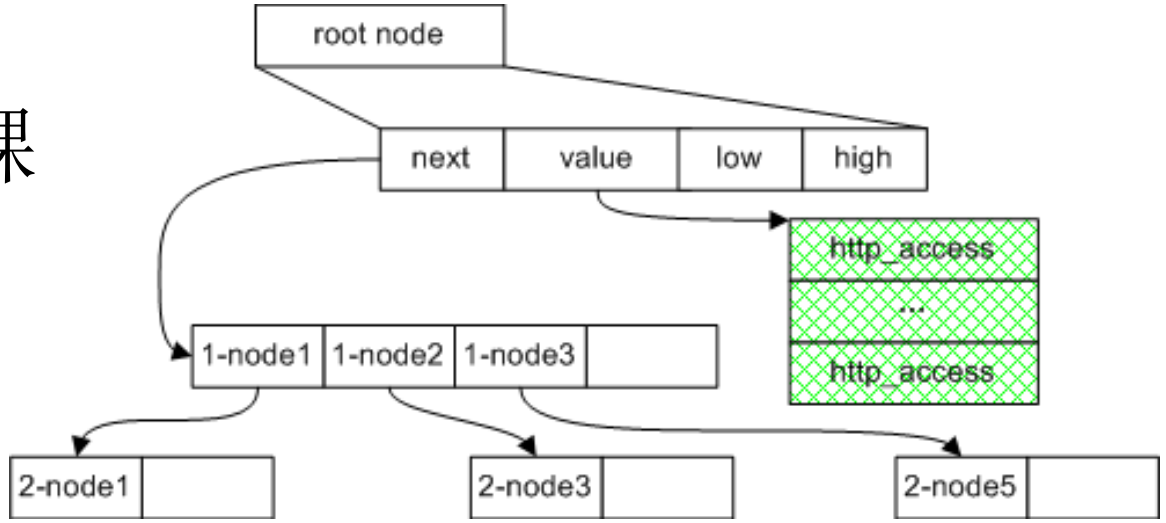


ACL优化

- squid根据http_access配置的顺序依次比较



- Swift将ACL按domain建立一棵Trie树

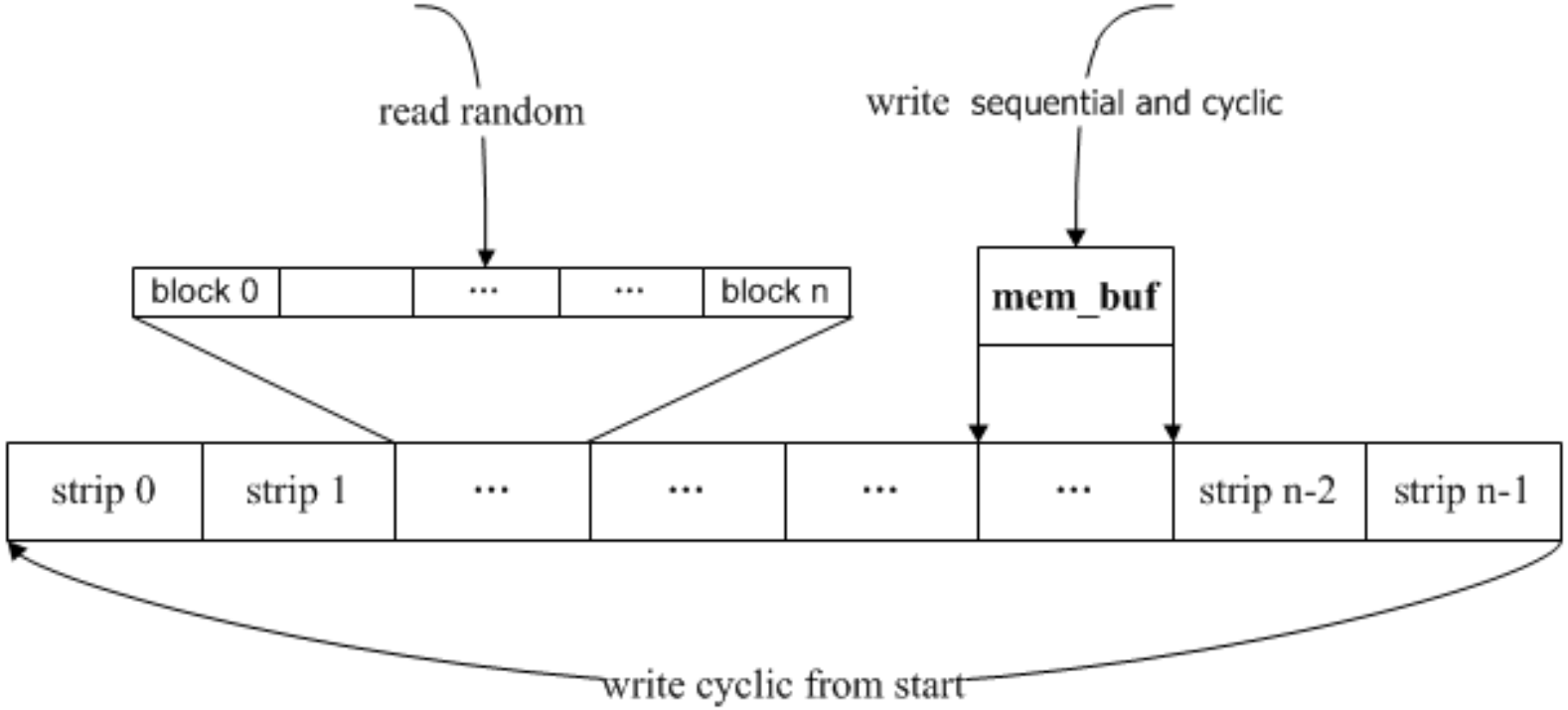


完美hash处理http header

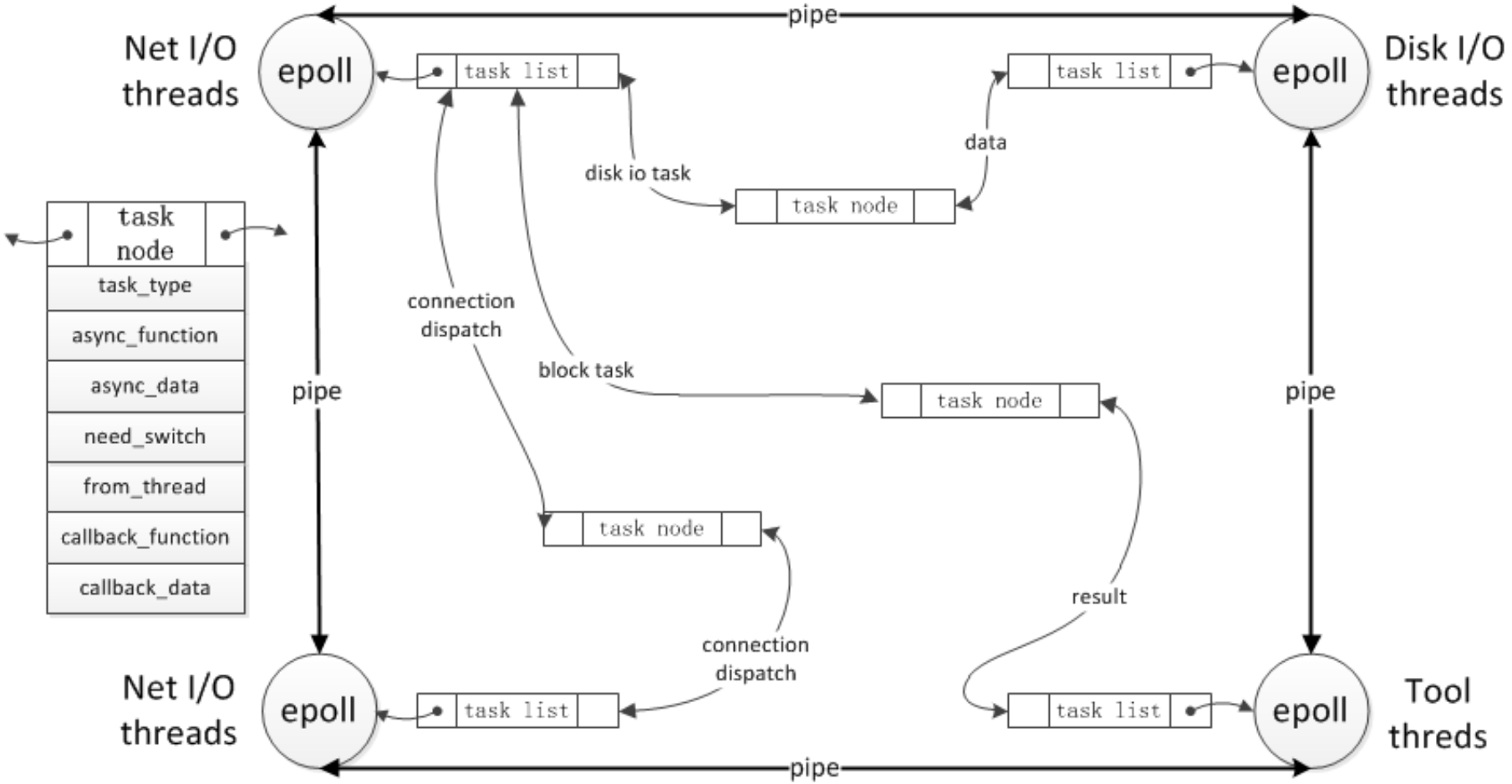
- 完美hash（ Perfect Hash Function ）
  - Hash table: key = value
  - PHF将key集合没有冲突地映射到一组整数
  - 查找key操作转换为索引整数表
- PHF场景&作用
  - 适合在key集合确定或不经常更新的情况
  - 主要作用是提高hash查找的速度

Store模块

- DIRECT IO写裸盘、绕过FS、不使用page cache
- 顺序写/随机读，stripe 8MB/block 512B
- mem_buf 在内存做写合并
- 省去open和close系统调用

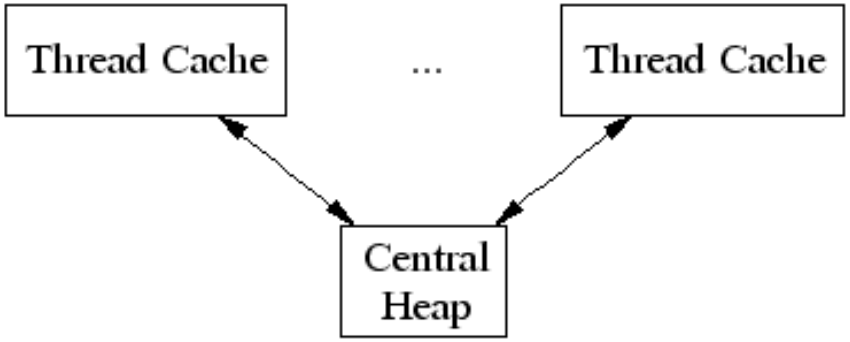


多线程任务交互模型

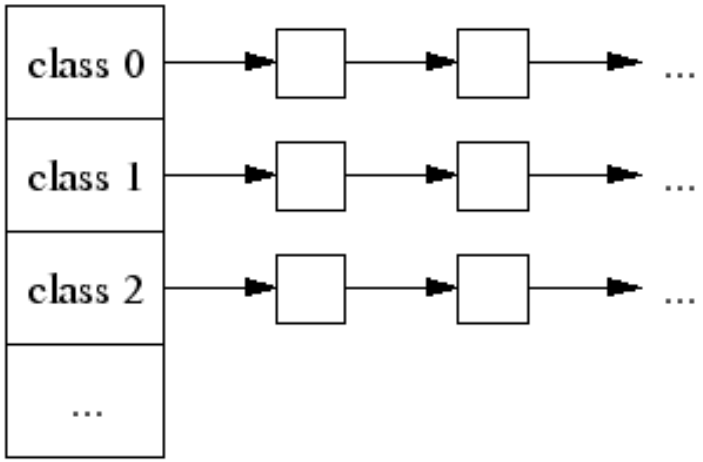

Net I/O threads
epoll
pipe
task list
epoll
Disk I/O threads
task list
data
disk io task
task node
task node
task_type
async_function
async_data
need_switch
from_thread
callback_function
callback_data
connection dispatch
block task
task node
pipe
pipe
task node
result
connection dispatch
Net I/O threads
epoll
task list
task list
epoll
Tool threds
pipe

TCmalloc管理内存

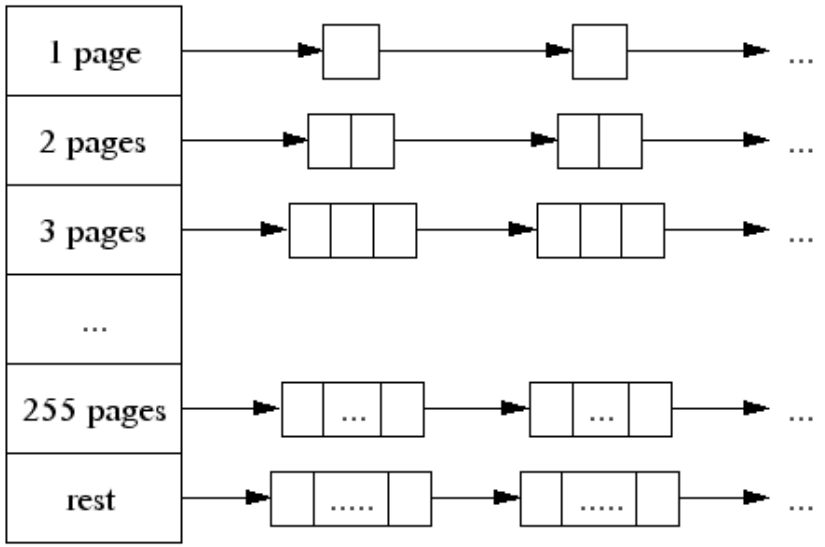
- thread free list
- central free list
- central heap



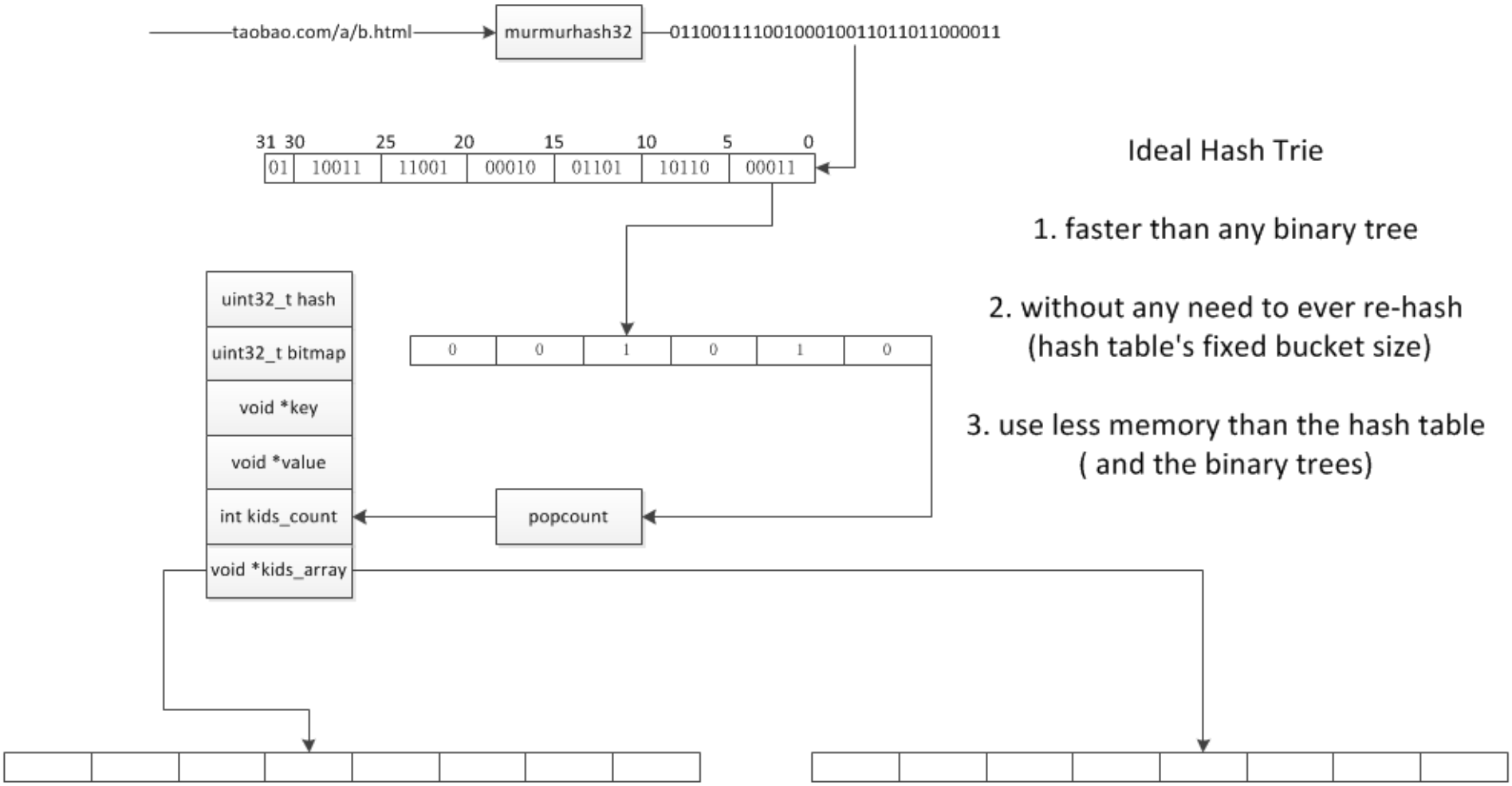
< 32KB



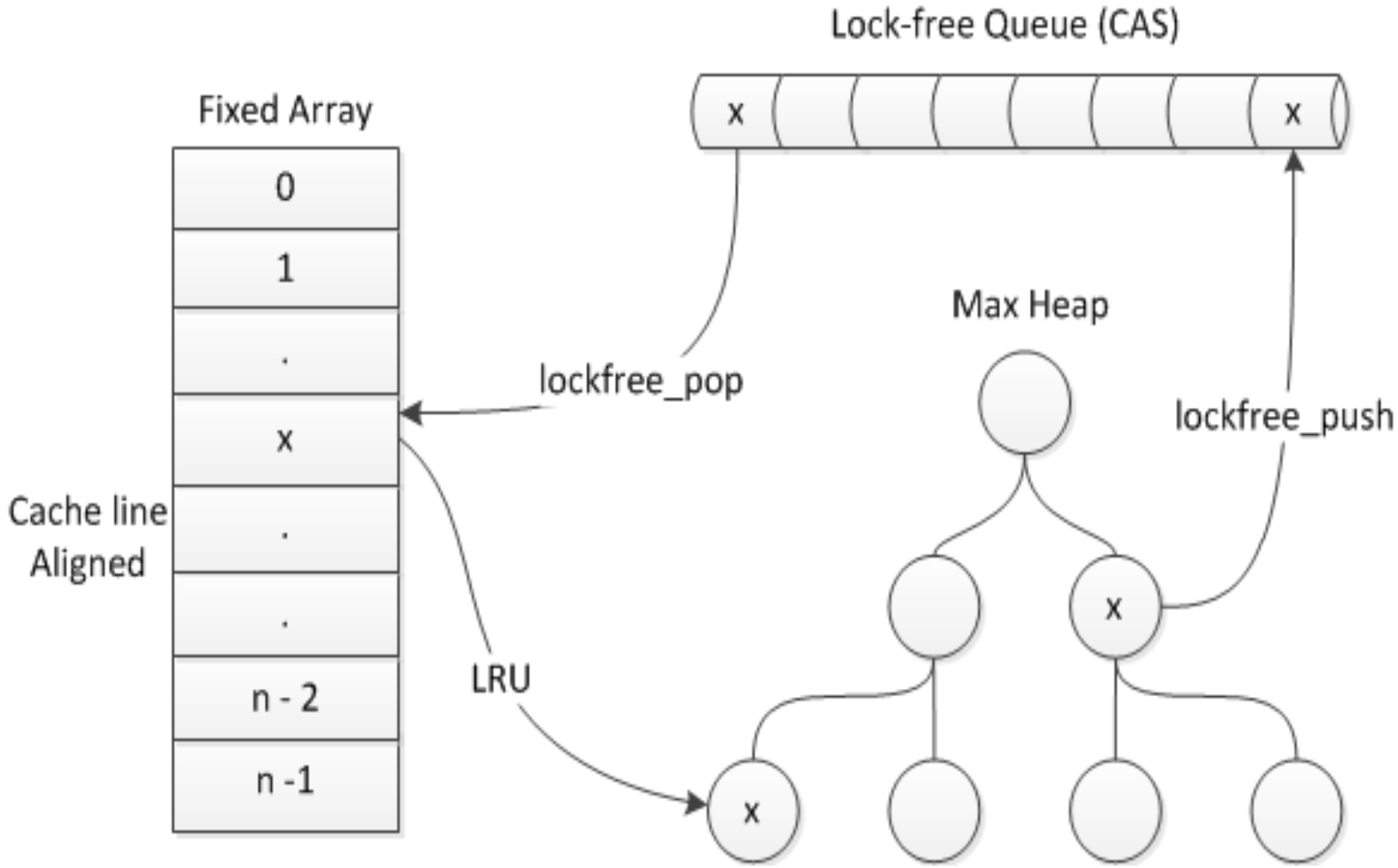
> 32KB



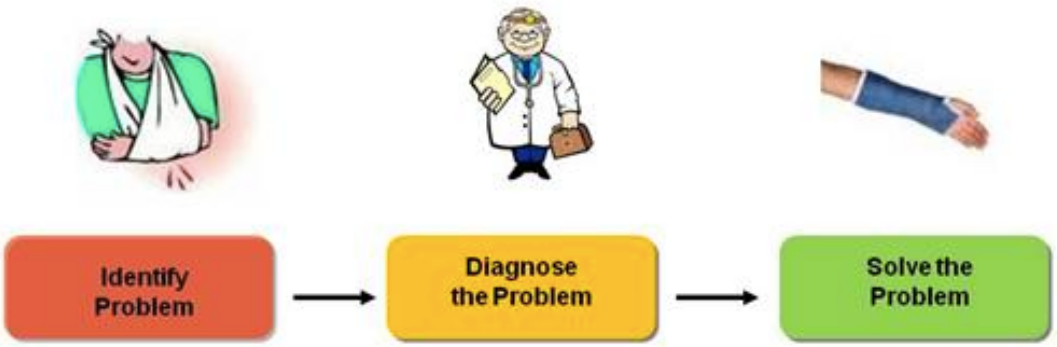
Hash Trie替换Hash table

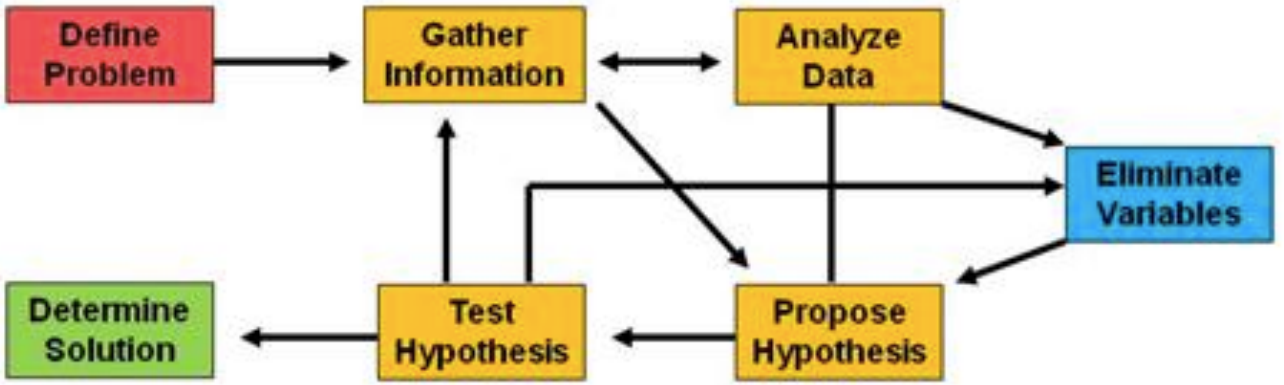


无锁LRU链表


Lock-free Queue (CAS)
x
x
Fixed Array
0
1
.
x
.
.
n - 2
n -1
Cache line
Aligned
lockfree_pop
LRU
Max Heap
lockfree_push
x
x

Troubleshooting





CPU分析工具

Linux	Solaris	Description
uptime	uptime	load averages
vmstat	vmstat	includes system-wide CPU averages
mpstat	mpstat	per-CPU statistics
sar	sar	historical statistics
ps	ps	process status
top	prstat	monitor per-process/thread CPU usage
pidstat	prstat	per-process/thread CPU breakdowns
time	ptime	time a command, with CPU breakdowns
DTrace, perf	DTrace	CPU profiling and tracing
perf	cpustat	CPU performance counter analysis

用/proc/$pid定位问题

- # top –cbp $pid
- # strace -cp $pid
- # ps -flp $pid
- # pstack $pid

- # cat /proc/$pid/**wchan**
- # cat /proc/$pid/**status**
- # cat /proc/$pid/**sched**
- # cat /proc/$pid/**schedstat**
- # cat /proc/$pid/**syscall**
- # cat /proc/$pid/**stack**

Mem分析工具

Linux	Solaris	Description
vmstat	vmstat	virtual and physical memory statistics
sar	sar	historical statistics
slabtop	::kmastat	kernel slab allocator statistics
ps	ps	process status
top	prstat	monitor per-process memory usage
pmap	pmap	process address space statistics
DTrace	DTrace	allocation tracing

/proc/meminfo 项的关系

- MemTotal = LowTotal + HighTotal
- MemFree = LowFree + HighFree
- Slab = SReclaimable + SUnreclaimable
- Active = Active(anon) + Active(file)
- Inactive = Inactive(anon) + Inactive(file)
- AnonPages + ?X? = Active(anon) + Inactive(anon)
- Buffers + Cached = Active(file) + Inactive(file) + ?X?
- AnonPages + Buffers + Cached = Active + Inactive
- SwapTotal = SwapFree + SwapCached

内存问题定位

- 内存泄露
  - # env HEAPPROFILE=/home/qianshi/dev/swift.hprof ./swift -f swift.conf
  - # pprof --pdf --base=swift.hprof.0001.heap . /swift swift.hprof.0002.heap > 1-2.pdf
- 内存写乱
  - # clang -O1 -g -fsanitize=address -fno-omit-frame-pointer example_UseAfterFree.cc

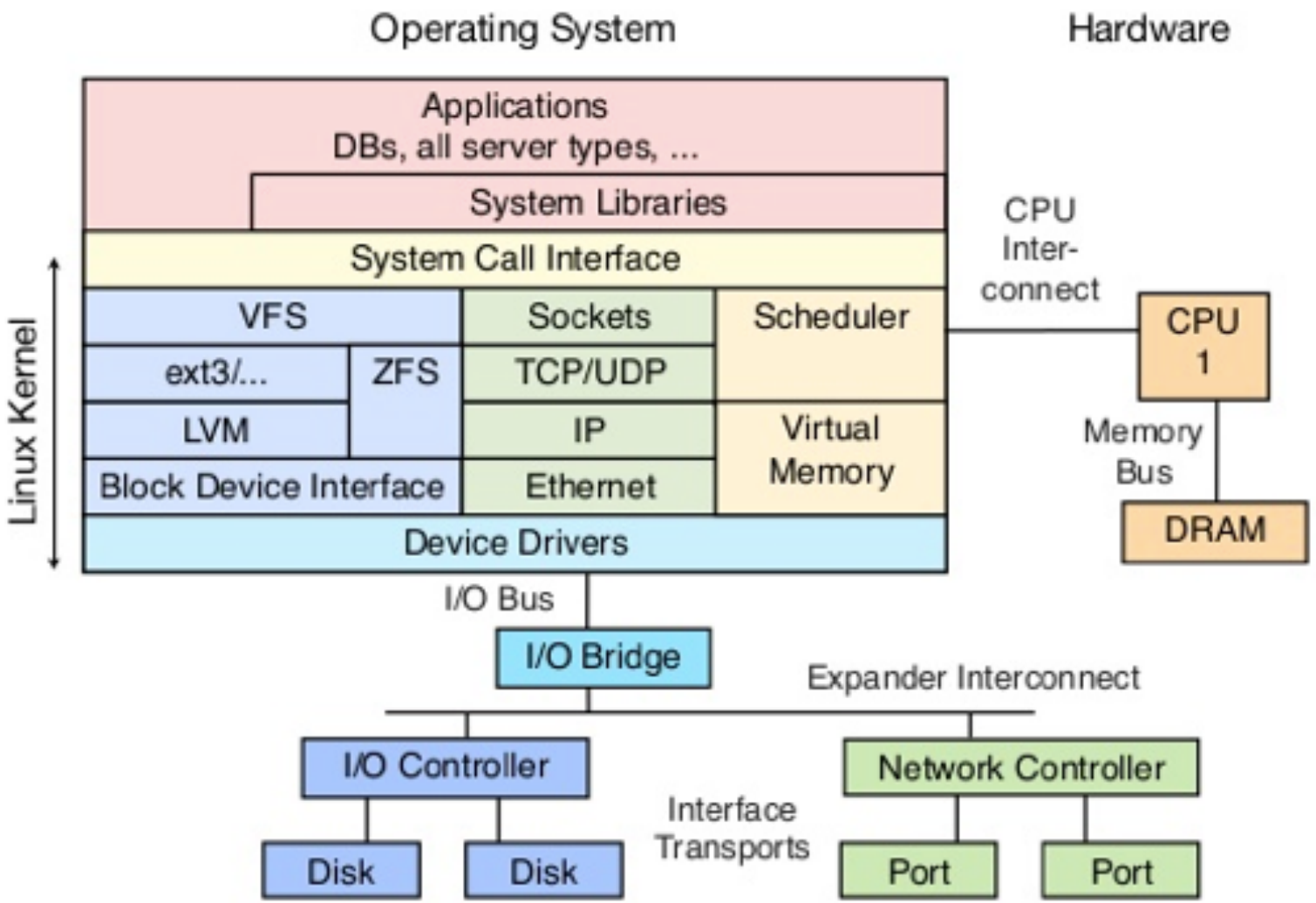
Disk分析工具

Linux	Solaris	Description
iostat	iostat	various per-disk statistics
sar	sar	historical disk statistics
pidstat, iotop	iotop	disk I/O usage by process
blktrace	iosnoop	disk I/O event tracing
DTrace	DTrace	custom static and dynamic tracing
MegaCli	MegaCli	LSI controller statistics
smartctl	smartctl	disk controller statistics

Network分析工具

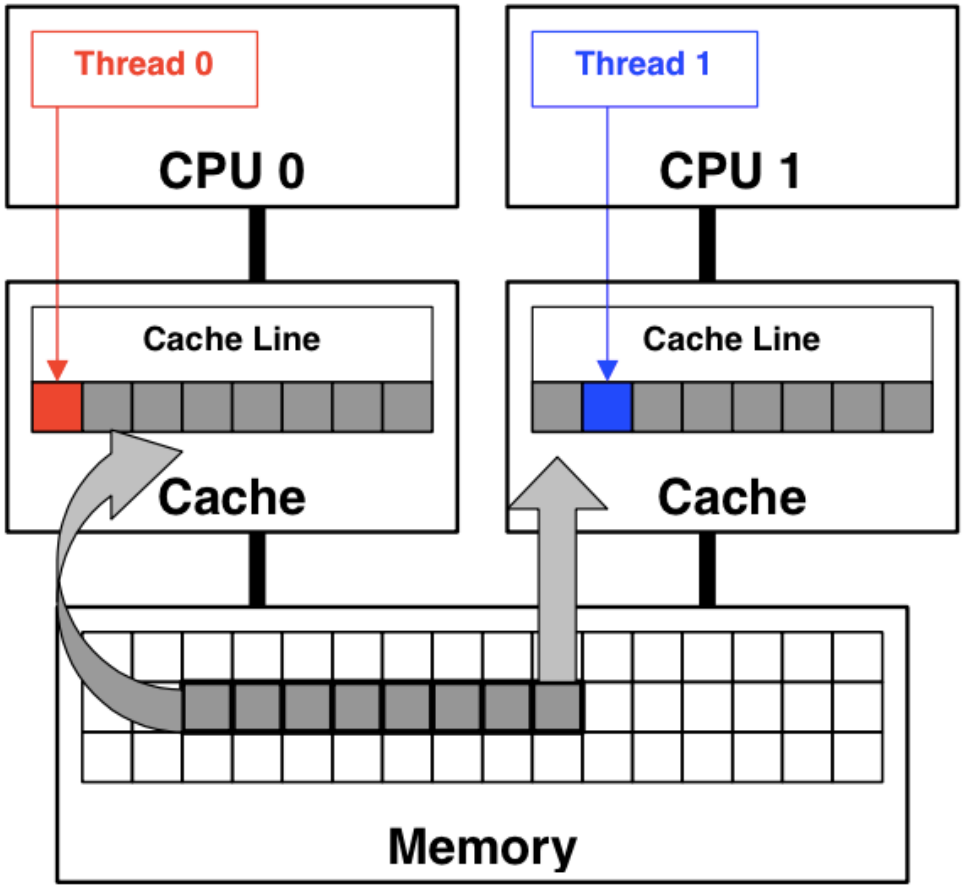
Linux	Solaris	Description
netstat	netstat	various network stack and interface statistics
sar	—	historical statistics
ifconfig	ifconfig	interface configuration
ip	dladm	network interface statistics
nicstat	nicstat	network interface throughput and utilization
ping	ping	test network connectivity
traceroute	traceroute	test network routes
pathchar	pathchar	determine network path characteristics
tcpdump	snoop/tcpdump	network packet sniffer
Wireshark	Wireshark	graphical network packet inspection
DTrace, perf	DTrace	TCP/IP stack tracing: connections, packets, drops, latency

Tuning


Operating System
Hardware
Applications
DBs, all server types, ...
System Libraries
System Call Interface
VFS
Sockets
Scheduler
ext3/...
ZFS
TCP/UDP
LVM
IP
Virtual Memory
Block Device Interface
Ethernet
Device Drivers
Linux Kernel
CPU Inter-connect
CPU 1
Memory Bus
DRAM
I/O Bus
I/O Bridge
Expander Interconnect
I/O Controller
Network Controller
Interface Transports
Disk
Disk
Port
Port

CPU tuning

- CPU亲和性
  - 提高cache命中率
  - 降低访问内存延迟
  - taskset –c –p $pid
- 避免false sharing
  - 编译器强制对齐
  - 填充结构体保证cache line对齐
  - 使用线程局部数据



Memory tuning

- 关掉SWAP
  - /proc/sys/vm/swappiness
  - swapoff –a

- OOM处理
  - /proc/$pid/oom_adj
  - /proc/sys/vm/overcommit_memory
  - /proc/sys/vm/overcommit_ratio

Disk tuning

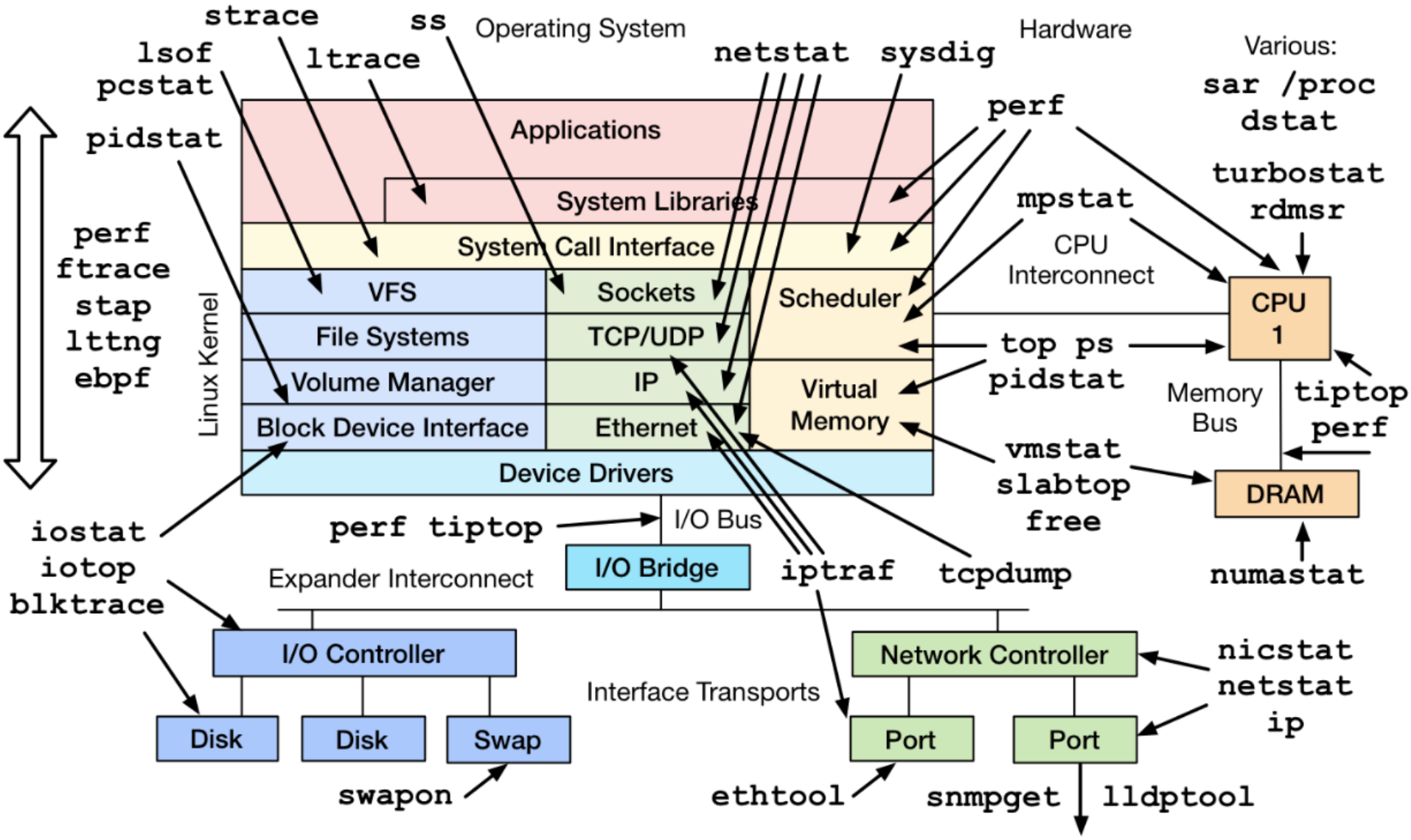
- Scheduler algorithm
  - echo deadline > /sys/block/<dev>/queue/scheduler
- IO request queue
  - echo 1024 > /sys/block/<dev>/queue/nr_requests

Network tuning

- Interrupts balance
  - /proc/irq/IRQ/smp_affinity
- Backlogs
  - net.core.netdev_max_backlog
  - net.core.somaxconn
  - net.ipv4.tcp_max_syn_backlog
- Reduce TCP overhead
  - net.ipv4.tcp_sack
  - net.ipv4.tcp_fack
- Reduce connection overhead
  - net.ipv4.tcp_fin_timeout
  - net.ipv4.tcp_tw_reuse
- Enable auto-tuning
  - net.ipv4.tcp_moderate_rcvbuf
  - net.ipv4.tcp_window_scaling

The tool isn't important – it's important to have a way to measure everything

-- Brendan Gregg

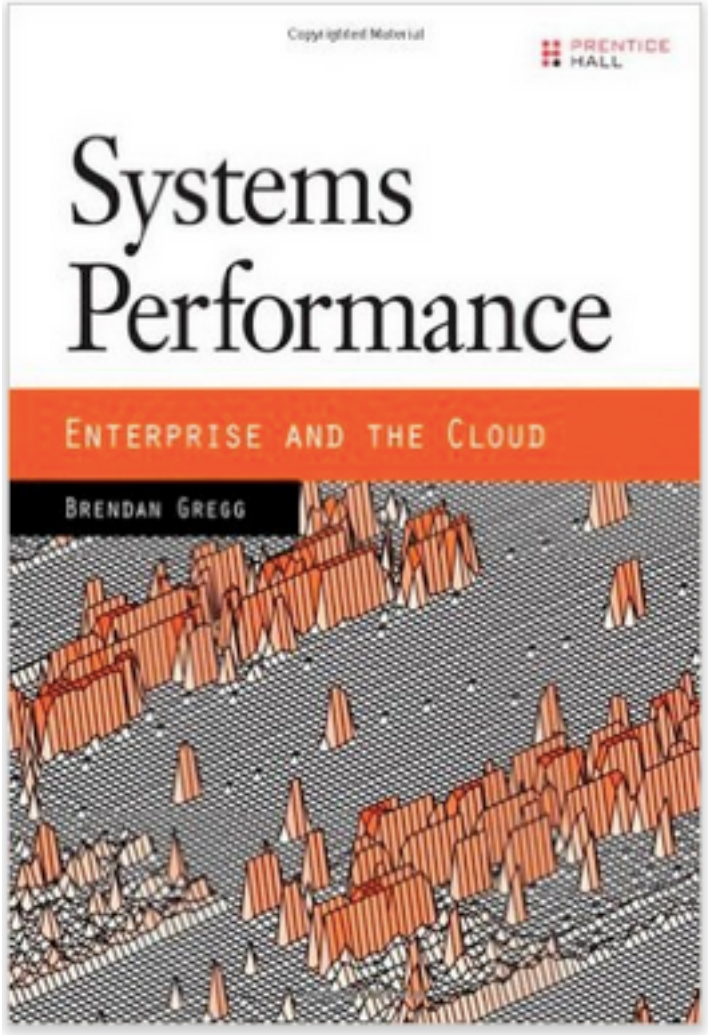

strace
ss
Operating System
Hardware
Various:
lsof
pcstat
ltrace
netstat
sysdig
sar /proc
dstat
perf
pidstat
turbostat
rdmsr
Applications
System Libraries
mpstat
perf
ftrace
stap
lttng
ebpf
System Call Interface
CPU
Interconnect
VFS
Sockets
Scheduler
CPU
1
Linux Kernel
File Systems
TCP/UDP
top ps
pidstat
Volume Manager
IP
Virtual
Memory
Memory
Bus
tiptop
perf
Block Device Interface
Ethernet
vmstat
slabtop
free
Device Drivers
DRAM
iostat
iotop
blktrace
perf tiptop
I/O Bus
I/O Bridge
iptraf
tcpdump
numastat
Expander Interconnect
I/O Controller
Network Controller
nicstat
netstat
ip
Interface Transports
Disk
Disk
Swap
Port
Port
swapon
ethtool
snmpget
lldptool

Architects look at thousands of buildings during their training, and study critiques of those buildings written by masters. In contrast, most software developers only ever get to know a handful of large programs well — usually programs they wrote themselves — and never study the great programs of history. As a result, they repeat one another's mistakes rather than building on one another's successes.

--The Architecture of Open Source Applications

推荐两本书


The Architecture of
Open Source Applications
Elegance, Evolution, and a Few Fearless Hacks
Edited by Amy Brown & Greg Wilson


Copyrighted Material
PRENTICE HALL
Systems
Performance
ENTERPRISE AND THE CLOUD
BRENDAN GREGG

reference

- http://tutorials.jenkov.com/software-architecture/computer-architecture.html
- http://exadat.co.uk/2015/01/29/cpus-memory-storage-and-database-engines-the-shape-of-things-to-come/
- http://mechanical-sympathy.blogspot.com/2013/02/cpu-cache-flushing-fallacy.html
- http://duartes.org/gustavo/blog/post/what-your-computer-does-while-you-wait/
- https://software.intel.com/en-us/articles/detecting-memory-bandwidth-saturation-in-threaded-applications
- https://software.intel.com/en-us/articles/optimizing-applications-for-numa
- https://www.thomas-krenn.com/en/wiki/Linux_Storage_Stack_Diagram
- http://www.mimuw.edu.pl/~lichota/09-10/Optymalizacja-open-source/Materialy/10%20-%20Dysk/gelato_ICE06apr_blktrace_brunelle_hp.pdf
- http://codecapsule.com/2014/02/12/coding-for-ssds-part-2-architecture-of-an-ssd-and-benchmarking/
- http://codecapsule.com/2014/02/12/coding-for-ssds-part-3-pages-blocks-and-the-flash-translation-layer/

- http://cd-docdb.fnal.gov/0019/001968/001/Linux-Pkt-Recv-Performance-Analysis-Final.pdf
- http://www.slideshare.net/hisaki/x86-hardware-for-packet-processing
- http://balodeamit.blogspot.com/2013/10/receive-side-scaling-and-receive-packet.html
- https://hpi.de/plattner/research/tools-methods-for-enterprise-systems-design-and-engineering.html
- http://sdepl.ucsd.edu/cgi-bin/yman2html?m=tcp&s=7
- http://zh.wikipedia.org/wiki/Trie
- http://en.wikipedia.org/wiki/Perfect_hash_function
- http://goog-perftools.sourceforge.net/doc/tcmalloc.html
- https://software.intel.com/en-us/articles/avoiding-and-identifying-false-sharing-among-threads
- https://www.certificationkits.com/cisco-certification/cisco-ccnp-tshoot-642-832-exam-study-center/cisco-ccnp-tshoot-troubleshooting-networks/
- http://brendangregg.com/books.html
- http://blog.tanelpoder.com/2013/02/21/peeking-into-linux-kernel-land-using-proc-filesystem-for-quickndirty-troubleshooting/
- http://goog-perftools.sourceforge.net/doc/heap_profiler.html
- http://clang.llvm.org/docs/AddressSanitizer.html
- http://brendangregg.com/linuxperf.html

淘宝千石

扫描上面的二维码，关注我吧