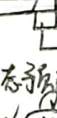
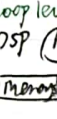
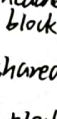
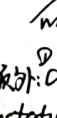
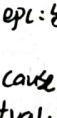
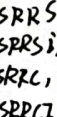



super scalar: 有2条及以上
 并行工作的流水级结构。
 并行流水线化支持Toma
 cycle not fixed in issue)
 = super pipeline: 增加流水
 级数: VLIW: 长指令,
 占用多条, 每阶段控制
 相应部件; 将长指令变为
 许多短指令。CRAY-1 RV64V
 = 并行研究: ILLIAC IV NV
 ISD: uniprocessor, 与并行
 指令并行; SIMD: 多种变
 体。Vector, AVX, GPU: M2SD
 并行; M2MD: ^{CUDA}利用task-
 level并行(指令级并行利用
 处理机任务则是多处理器
 MA: symmetric (shared
 memory) multiprocessor (SMP)
 centralized shared mem-
 ory multiprocessor, 访问时间
 一样, 也称为 tightly coupled
 memory MUMA is cell-
 distributed shared memory
 multiprocessor, 访问时间和
 延迟均不同, 故
 设计时要尽量降低延迟地
 访问区, misses are seen in the
 same order 一致。
 cache coherence 一致。
 与列最新数据。保证
 计算机读一致: cache
 consistency 一致。
 = cache coherence 一致。
 在分布系统中一样。
 noop for B/Mp. Directory
 Bsp, 对Dir的。
 read: 一个节点有, cache
 都最新; uncached
 在cache中, 内存最新
 一个node有最新,
 无最新。

Superscalar: static (scheduling) in dynamic (dynamic) with speculative (dynamic) with
 VLIW: static (hardware), software (compiler)
 Superscalar:  VLIW: 
 动态标度:  SMP: 
 DLA: 
 Loop level: 
 PSP (Multicore MP): 
 Memory:
 Interconnection network:
 对 directory:
 uncached:
 shared block:
 block:
 exclusive:
 Date write back:
 write miss:
 额外:
 mstatus:
 保存进入异常之前特权模式的MP:
 mtime:
 mepc:
 mcause:
 mtimeval:
 CSRRW rd, csr, rs1
 CSRRW rd, csr, imm
 CSRRS rd, csr, rs1 或 rd, csr, imm
 CSRRS rd, csr, imm
 CSRRC rd, csr, rs1 再根据 rd 对 csr
 CSRRCT rd, csr, imm (根据 rd 对 csr)
 ② 硬件原语: 常用 CAS (内存位置 A 的当前值 A 新值 B 若为 A 则
 否则为 B; 1: 成功, 0: 失败)

order 顺序
无 speculation
speculation) 乱序有 spec
ection, 旁有内存及并行器
er pipe line
N
P₂ P₃
re cache
sred cache
Main Mem
loop parallel
ular
k
sha node
ing node
t
R的 block
n directory, memory
cached share set
主拥有, 互斥
包括开集中断的 M2
model CPU host GPU device, unit all forms
of GPU as Cum thread program model is Sing
the instruction Multiple thread "GPU 线程
ramp housekeeping 多线程管理而非 OS 管
GPU mem shared by all GPUs @ thread block
block with grid eliminate main GPU A thread co
并行编程编程师 GPU physical mem is uni
formly shared by all processes
@ num: All cpts share a
uniform address space
local cache of mem
main mem, cache 缓存
可以跟软件完全
模拟, 可知操作
2. 工业成功: 实现 store
SCLL: 可以构建更
但层法, 与 GPU 相比
读取和写入方式, 解
决多个看到致的不
B 阻塞值) 恰当竞争
Horizontal
ractor N20
交互 20
核 32 核 16
Vertical (N2)
mem mem
GPU 核 16 分
组, N=32M, 组内核同
组, 每 16 核 16 核, 时
并行冲突 @ 核未 let out
是更多的并行部件来增大并行
不可能 link; interleaving mixing;
用 multi-processor system 设计
技术里有并行并行, 取 16 核
N=1 argument vector 0
ray processor N repeated
核, Interconnect to Form
a array; control unit, 同指令的
并行核; 有时 parallel processor
@ distributed centralized shared mem
用直联网 N 直联, 是 AG
P2M 和 P2-- 对应; 后者 NTP 对 N
Item, 通过 N 直联, 而 the connect
ion architecture for parallel comput
是连接核, 连接核直联网
路和上层通信, 编译器 CPU,
mem interface, link (bri),
switch node), 均为 static
indynamic @ single
stage fabric 内部,
P2M 和 shuffle 1/3
@ 1/2 mod N 2/3
Grid 格网路由接收
直联网 switch bus
rich control method and
Topology 拓扑 GPU
heterogeneous elec
model CPU host GPU device, unit all forms
of GPU as Cum thread program model is Sing
the instruction Multiple thread "GPU 线程
ramp housekeeping 多线程管理而非 OS 管
GPU mem shared by all GPUs @ thread block
block with grid eliminate main GPU A thread co
并行编程编程师 GPU physical mem is uni
formly shared by all processes
@ num: All cpts share a
uniform address space
local cache of mem
main mem, cache 缓存
可以跟软件完全
模拟, 可知操作
2. 工业成功: 实现 store
SCLL: 可以构建更
但层法, 与 GPU 相比
读取和写入方式, 解
决多个看到致的不
B 阻塞值) 恰当竞争
Horizontal
ractor N20
交互 20
核 32 核 16
Vertical (N2)
mem mem
GPU 核 16 分
组, N=32M, 组内核同
组, 每 16 核 16 核, 时
并行冲突 @ 核未 let out
是更多的并行部件来增大并行
不可能 link; interleaving mixing;
用 multi-processor system 设计
技术里有并行并行, 取 16 核
N=1 argument vector 0
ray processor N repeated
核, Interconnect to Form
a array; control unit, 同指令的
并行核; 有时 parallel processor
@ distributed centralized shared mem
用直联网 N 直联, 是 AG
P2M 和 P2-- 对应; 后者 NTP 对 N
Item, 通过 N 直联, 而 the connect
ion architecture for parallel comput
是连接核, 连接核直联网
路和上层通信, 编译器 CPU,
mem interface, link (bri),
switch node), 均为 static
indynamic @ single
stage fabric 内部,
P2M 和 shuffle 1/3
@ 1/2 mod N 2/3
Grid 格网路由接收
直联网 switch bus
rich control method and
Topology 拓扑 GPU
heterogeneous elec
model CPU host GPU device, unit all forms
of GPU as Cum thread program model is Sing
the instruction Multiple thread "GPU 线程
ramp housekeeping 多线程管理而非 OS 管
GPU mem shared by all GPUs @ thread block
block with grid eliminate main GPU A thread co
并行编程编程师 GPU physical mem is uni
formly shared by all processes
@ num: All cpts share a
uniform address space
local cache of mem
main mem, cache 缓存
可以跟软件完全
模拟, 可知操作
2. 工业成功: 实现 store
SCLL: 可以构建更
但层法, 与 GPU 相比
读取和写入方式, 解
决多个看到致的不
B 阻塞值) 恰当竞争
Horizontal
ractor N20
交互 20
核 32 核 16
Vertical (N2)
mem mem
GPU 核 16 分
组, N=32M, 组内核同
组, 每 16 核 16 核, 时
并行冲突 @ 核未 let out
是更多的并行部件来增大并行
不可能 link; interleaving mixing;
用 multi-processor system 设计
技术里有并行并行, 取 16 核
N=1 argument vector 0
ray processor N repeated
核, Interconnect to Form
a array; control unit, 同指令的
并行核; 有时 parallel processor
@ distributed centralized shared mem
用直联网 N 直联, 是 AG
P2M 和 P2-- 对应; 后者 NTP 对 N
Item, 通过 N 直联, 而 the connect
ion architecture for parallel comput
是连接核, 连接核直联网
路和上层通信, 编译器 CPU,
mem interface, link (bri),
switch node), 均为 static
indynamic @ single
stage fabric 内部,
P2M 和 shuffle 1/3
@ 1/2 mod N 2/3
Grid 格网路由接收
直联网 switch bus
rich control method and
Topology 拓扑 GPU
heterogeneous elec
model CPU host GPU device, unit all forms
of GPU as Cum thread program model is Sing
the instruction Multiple thread "GPU 线程
ramp housekeeping 多线程管理而非 OS 管
GPU mem shared by all GPUs @ thread block
block with grid eliminate main GPU A thread co
并行编程编程师 GPU physical mem is uni
formly shared by all processes
@ num: All cpts share a
uniform address space
local cache of mem
main mem, cache 缓存
可以跟软件完全
模拟, 可知操作
2. 工业成功: 实现 store
SCLL: 可以构建更
但层法, 与 GPU 相比
读取和写入方式, 解
决多个看到致的不
B 阻塞值) 恰当竞争
Horizontal
ractor N20
交互 20
核 32 核 16
Vertical (N2)
mem mem
GPU 核 16 分
组, N=32M, 组内核同
组, 每 16 核 16 核, 时
并行冲突 @ 核未 let out
是更多的并行部件来增大并行
不可能 link; interleaving mixing;
用 multi-processor system 设计
技术里有并行并行, 取 16 核
N=1 argument vector 0
ray processor N repeated
核, Interconnect to Form
a array; control unit, 同指令的
并行核; 有时 parallel processor
@ distributed centralized shared mem
用直联网 N 直联, 是 AG
P2M 和 P2-- 对应; 后者 NTP 对 N
Item, 通过 N 直联, 而 the connect
ion architecture for parallel comput
是连接核, 连接核直联网
路和上层通信, 编译器 CPU,
mem interface, link (bri),
switch node), 均为 static
indynamic @ single
stage fabric 内部,
P2M 和 shuffle 1/3
@ 1/2 mod N 2/3
Grid 格网路由接收
直联网 switch bus
rich control method and
Topology 拓扑 GPU
heterogeneous elec
model CPU host GPU device, unit all forms
of GPU as Cum thread program model is Sing
the instruction Multiple thread "GPU 线程
ramp housekeeping 多线程管理而非 OS 管
GPU mem shared by all GPUs @ thread block
block with grid eliminate main GPU A thread co
并行编程编程师 GPU physical mem is uni
formly shared by all processes
@ num: All cpts share a
uniform address space
local cache of mem
main mem, cache 缓存
可以跟软件完全
模拟, 可知操作
2. 工业成功: 实现 store
SCLL: 可以构建更
但层法, 与 GPU 相比
读取和写入方式, 解
决多个看到致的不
B 阻塞值) 恰当竞争
Horizontal
ractor N20
交互 20
核 32 核 16
Vertical (N2)
mem mem
GPU 核 16 分
组, N=32M, 组内核同
组, 每 16 核 16 核, 时
并行冲突 @ 核未 let out
是更多的并行部件来增大并行
不可能 link; interleaving mixing;
用 multi-processor system 设计
技术里有并行并行, 取 16 核
N=1 argument vector 0
ray processor N repeated
核, Interconnect to Form
a array; control unit, 同指令的
并行核; 有时 parallel processor
@ distributed centralized shared mem
用直联网 N 直联, 是 AG
P2M 和 P2-- 对应; 后者 NTP 对 N
Item, 通过 N 直联, 而 the connect
ion architecture for parallel comput
是连接核, 连接核直联网
路和上层通信, 编译器 CPU,
mem interface, link (bri),
switch node), 均为 static
indynamic @ single
stage fabric 内部,
P2M 和 shuffle 1/3
@ 1/2 mod N 2/3
Grid 格网路由接收
直联网 switch bus
rich control method and
Topology 拓扑 GPU
heterogeneous elec
model CPU host GPU device, unit all forms
of GPU as Cum thread program model is Sing
the instruction Multiple thread "GPU 线程
ramp housekeeping 多线程管理而非 OS 管
GPU mem shared by all GPUs @ thread block
block with grid eliminate main GPU A thread co
并行编程编程师 GPU physical mem is uni
formly shared by all processes
@ num: All cpts share a
uniform address space
local cache of mem
main mem, cache 缓存
可以跟软件完全
模拟, 可知操作
2. 工业成功: 实现 store
SCLL: 可以构建更
但层法, 与 GPU 相比
读取和写入方式, 解
决多个看到致的不
B 阻塞值) 恰当竞争
Horizontal
ractor N20
交互 20
核 32 核 16
Vertical (N2)
mem mem
GPU 核 16 分
组, N=32M, 组内核同
组, 每 16 核 16 核, 时
并行冲突 @ 核未 let out
是更多的并行部件来增大并行
不可能 link; interleaving mixing;
用 multi-processor system 设计
技术里有并行并行, 取 16 核
N=1 argument vector 0
ray processor N repeated
核, Interconnect to Form
a array; control unit, 同指令的
并行核; 有时 parallel processor
@ distributed centralized shared mem
用直联网 N 直联, 是 AG
P2M 和 P2-- 对应; 后者 NTP 对 N
Item, 通过 N 直联, 而 the connect
ion architecture for parallel comput
是连接核, 连接核直联网
路和上层通信, 编译器 CPU,
mem interface, link (bri),
switch node), 均为 static
indynamic @ single
stage fabric 内部,
P2M 和 shuffle 1/3
@ 1/2 mod N 2/3
Grid 格网路由接收
直联网 switch bus
rich control method and
Topology 拓扑 GPU
heterogeneous elec
model CPU host GPU device, unit all forms
of GPU as Cum thread program model is Sing
the instruction Multiple thread "GPU 线程
ramp housekeeping 多线程管理而非 OS 管
GPU mem shared by all GPUs @ thread block
block with grid eliminate main GPU A thread co
并行编程编程师 GPU physical mem is uni
formly shared by all processes
@ num: All cpts share a
uniform address space
local cache of mem
main mem, cache 缓存
可以跟软件完全
模拟, 可知操作
2. 工业成功: 实现 store
SCLL: 可以构建更
但层法, 与 GPU 相比
读取和写入方式, 解
决多个看到致的不
B 阻塞值) 恰当竞争
Horizontal
ractor N20
交互 20
核 32 核 16
Vertical (N2)
mem mem
GPU 核 16 分
组, N=32M, 组内核同
组, 每 16 核 16 核, 时
并行冲突 @ 核未 let out
是更多的并行部件来增大并行
不可能 link; interleaving mixing;
用 multi-processor system 设计
技术里有并行并行, 取 16 核
N=1 argument vector 0
ray processor N repeated
核, Interconnect to Form
a array; control unit, 同指令的
并行核; 有时 parallel processor
@ distributed centralized shared mem
用直联网 N 直联, 是 AG
P2M 和 P2-- 对应; 后者 NTP 对 N
Item, 通过 N 直联, 而 the connect
ion architecture for parallel comput
是连接核, 连接核直联网
路和上层通信, 编译器 CPU,
mem interface, link (bri),
switch node), 均为 static
indynamic @ single
stage fabric 内部,
P2M 和 shuffle 1/3
@ 1/2 mod N 2/3
Grid 格网路由接收
直联网 switch bus
rich control method and
Topology 拓扑 GPU
heterogeneous elec
model CPU host GPU device, unit all forms
of GPU as Cum thread program model is Sing
the instruction Multiple thread "GPU 线程
ramp housekeeping 多线程管理而非 OS 管
GPU mem shared by all GPUs @ thread block
block with grid eliminate main GPU A thread co
并行编程编程师 GPU physical mem is uni
formly shared

[illegible]

hit time 到 1 cycle
⑤ 所有指令 100%
指令中的 load/store
为 50% (流水线)
和 use resource 的时间间隔
Initiation level: 两个
失操作间等待的周期数
流水 add, mul, div 都是 1 个
的延迟 (25 个周期)
⑥ 为流水线设计指令集
时 (改变时间) 变; X86 的
cached addressing modes;
寄存器地址; X 修改自身代
码; 改变 implicitly setting
CCs in instruction
write buffer: 缓冲 stall
但不能完全解决 stall
SAMD, 32m Power7
superscalar 第 1-8 个
load/store 在 1 个 integer
operation, 其它算 float
可以增加 stage Time
for cycle 流出一个
[temporal] locality 指 need
the requested word again
spatial locality 指 need
other data in the block
space
[latency] -ve time to
retrieve the first word
of block; Bandwidth: the
time to retrieve the rest
of this block
如单 port code, port 的
number of And gate;
port number of And gate;
C_p 个 Flap-flap
可通过 code schedu
ling 来避免 stalls
通信点, 凡 MIP nodes to
discrete kernel reside
MPP 的 网络没有 common
network; core local
area network
以块 ② 问一个地址在 cache
是否 index + offset 所以 3
cache 多的位就长加
③ 块为等值算读与加
5/1 物位数 = 前地址 + 1 的位数
④ 56MB 缓存 = 2²⁶ 地址址
⑤ 的延迟: 100 指令, 5% miss
每指令 1.2 次访问, 平均指
引时间 = $4 \times 1.25 + 1.2 \times 5\%$
⑥ 5s = 10.1ms ⑦ TLB 命中
一次消耗 100 pages 命中
⑧ 虚存对后程序为反
透明 对系统透明看到
⑨ TLB 由组相联支持能黑
成, 不在内存也存 SRAM 组
cache 由 SRAM 而成 ⑩ TLB
也可能给个表, 此的取有址
和 index ⑪ 虚存看 TLB =
若有看 cache, 没有内存
否找 → cache 有直接出
内存内存找 ⑫ 算 cache
量要注意组相联时
以组号
coherency: dependability, sca
lability, efficient throughput
③ addressing mode: re
fer indirect, displacement
immediate ④ little endian
ntel, big, 32m, M, 16
⑤ 16b/32b separate file
point registers ⑥ caller-
ing: 保存所有寄存器用
r call, then invoke; callee-
3 invoke, 再 callee save re
gister (produced inst set, com
⑦ 和找本地址 = 79 条
的指数 $(k+n)/ot$
如重比 = 不取用本地址
用流水时间
⑧ 流水线要 restore local
offending inst. when 需
⑨ Latency: 在利生



