GPU TECHNOLOGY CONFERENCE

GPUDirect:集成具有网络接口的 GPU

DAVIDE ROSSETTI,SW 计算团队

GPUDirect 系列1

- GPUDirect Shared GPU-Sysmem 用于节点间复制优化
- GPUDirect P2P 用于节点内、加速 GPU-GPU memcpy
- GPUDirect P2P，用于节点内、GPU 间 LD/ST 访问
- GPUDirect RDMA2用于节点间复制优化

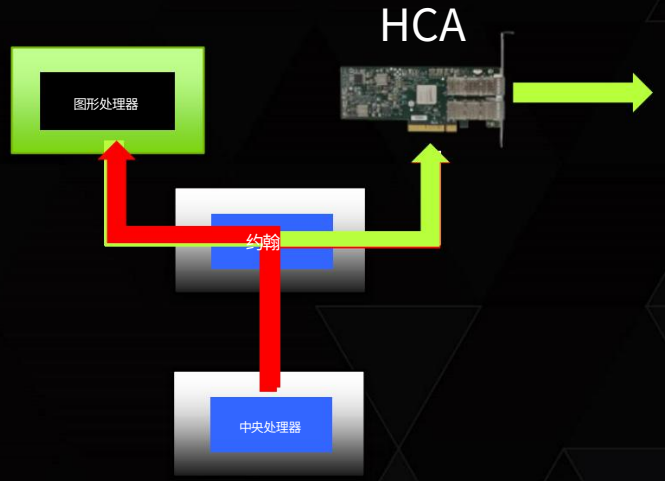
[1] 开发者信息：https: //developer.nvidia.com/gpudirect

[2] http://docs.nvidia.com/cuda/gpudirect-rdma

GPUDIRECT RDMA 功能 & 局限性

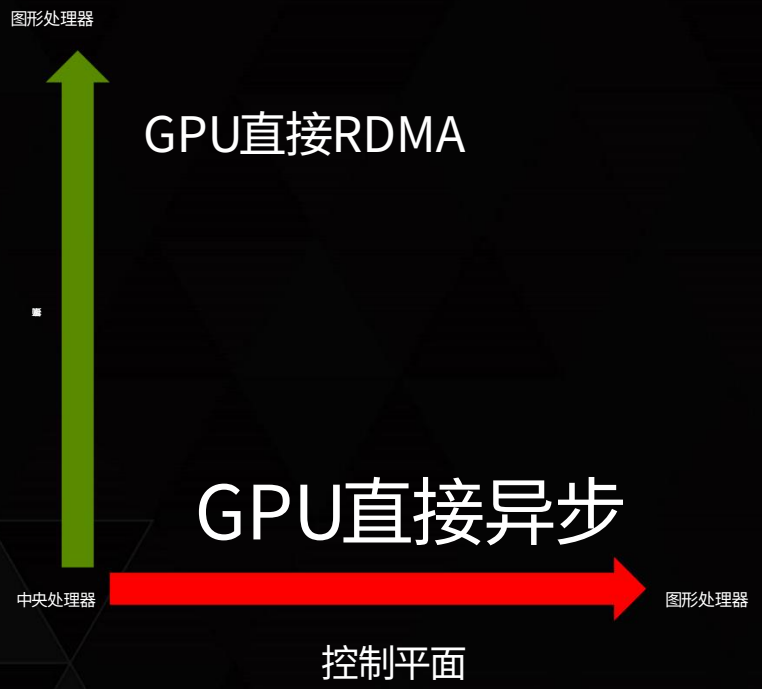
- GPU直接RDMA
  - 直接 HCA 访问 GPU 内存
- CPU仍驱动计算+通信
  - 需要快速CPU
  - 影响:功耗、延迟、TCO
  - 风险:规模有限……

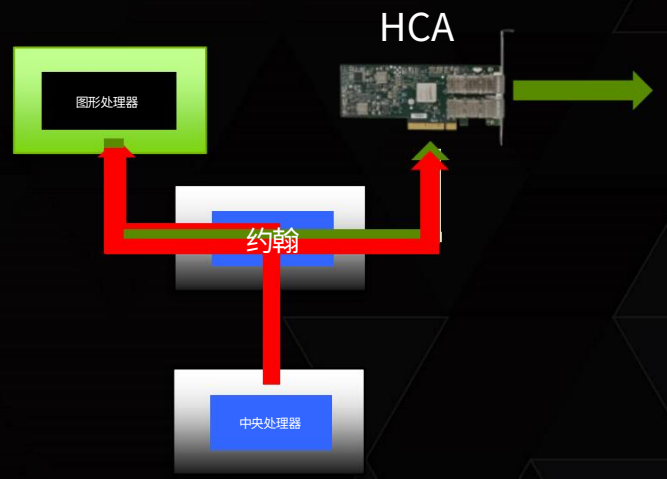
移动数据



HCA直接绕过系统发出的GPU内存通信任务

遇见下一件事情

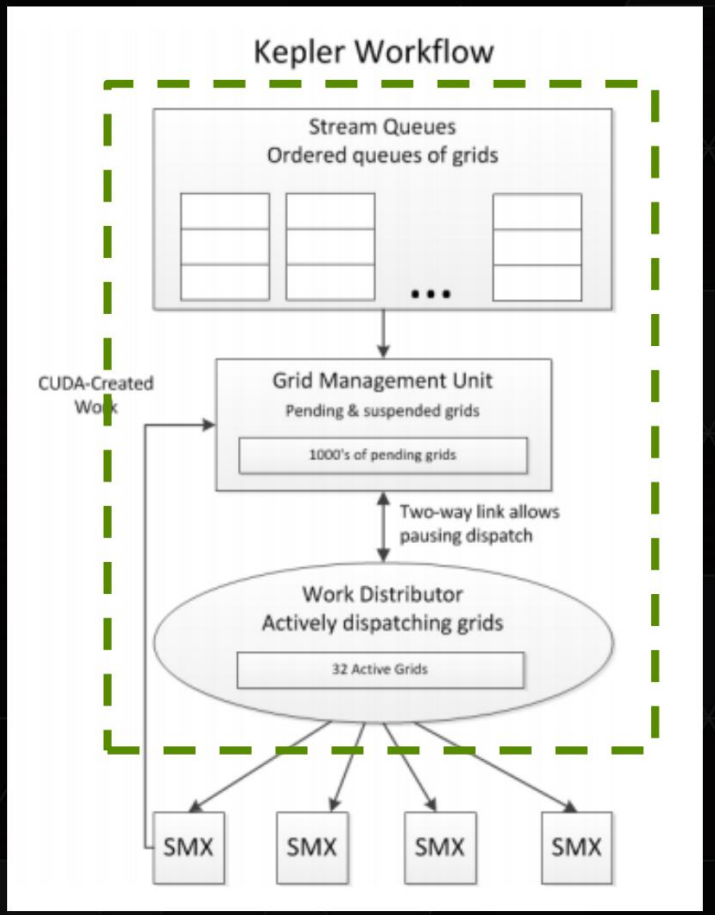




CPU 在 GPU 上准备计算和
通信任务并对其进行排队
GPU触发HCA上的通信
HCA直接访问GPU内存

CPU 关闭关键路径

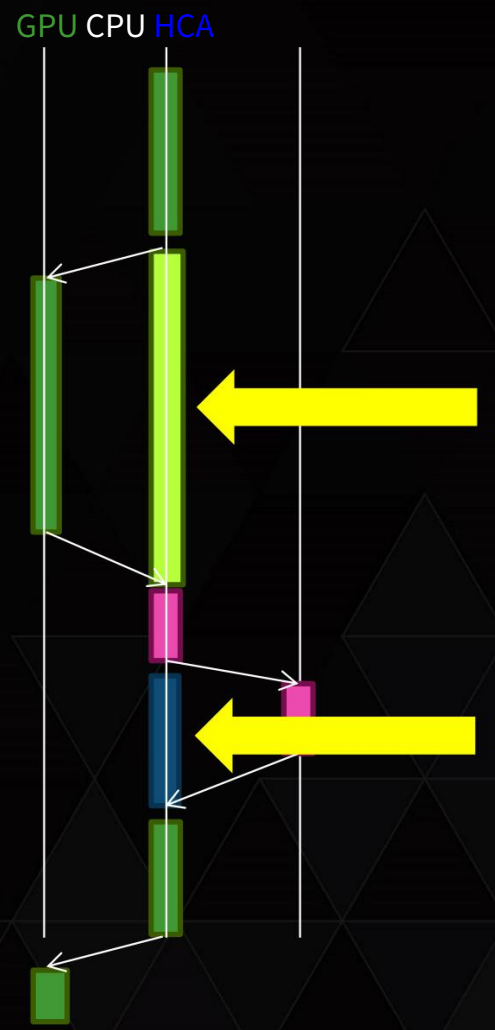
- CPU 准备工作计划难以并行、分支密集
- GPU 编排流程
  - 在优化调度单元上运行
  - 同一调度 GPU 工作
  - 现在还可以调度网络通信



内核+发送

正常流量

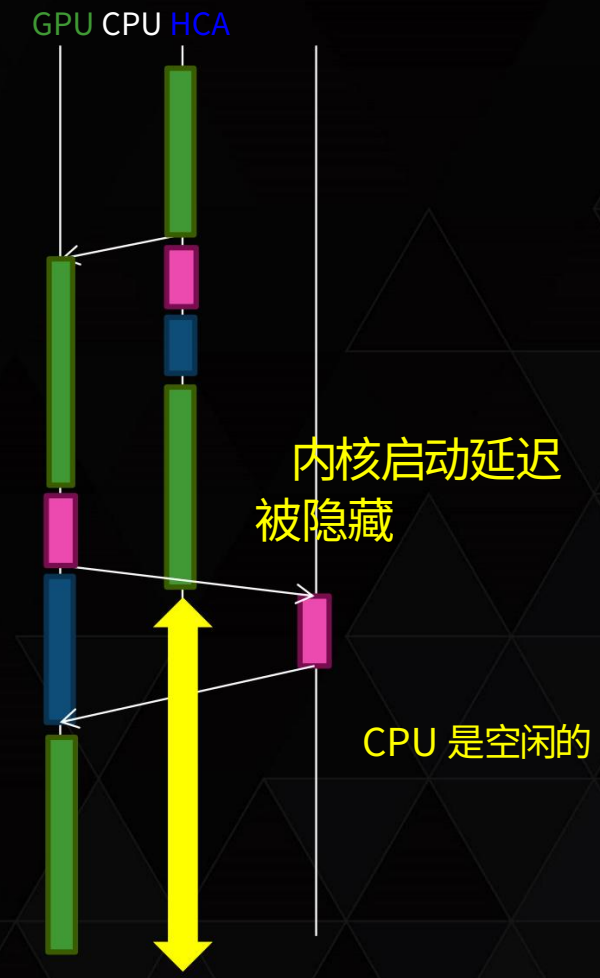
```
a_kernel<<<...,stream>>>(buf);
cudaStreamSynchronize(流);
ibv_post_send(buf);
while (!done) ibv_poll_cq(txcq);
b_kernel<<<...,stream>>>(buf);
```



100% CPU utilization
Limited scaling!

内核+发送
GPU直接异步

```
a_kernel<<<...,stream>>>(buf);
gds_stream_queue_send(流,qp,buf);
gds_stream_wait_cq(流,txcq);
b_kernel<<...,stream>>(buf);
```

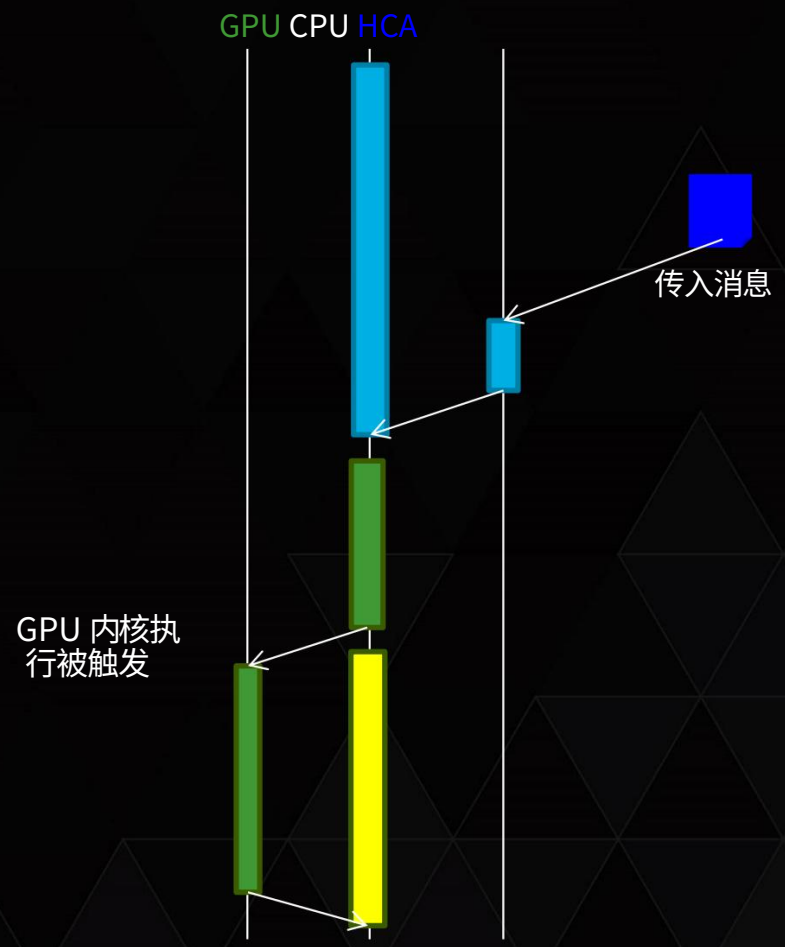


No CPU in critical path!
Improve Scaling!

接收+内核

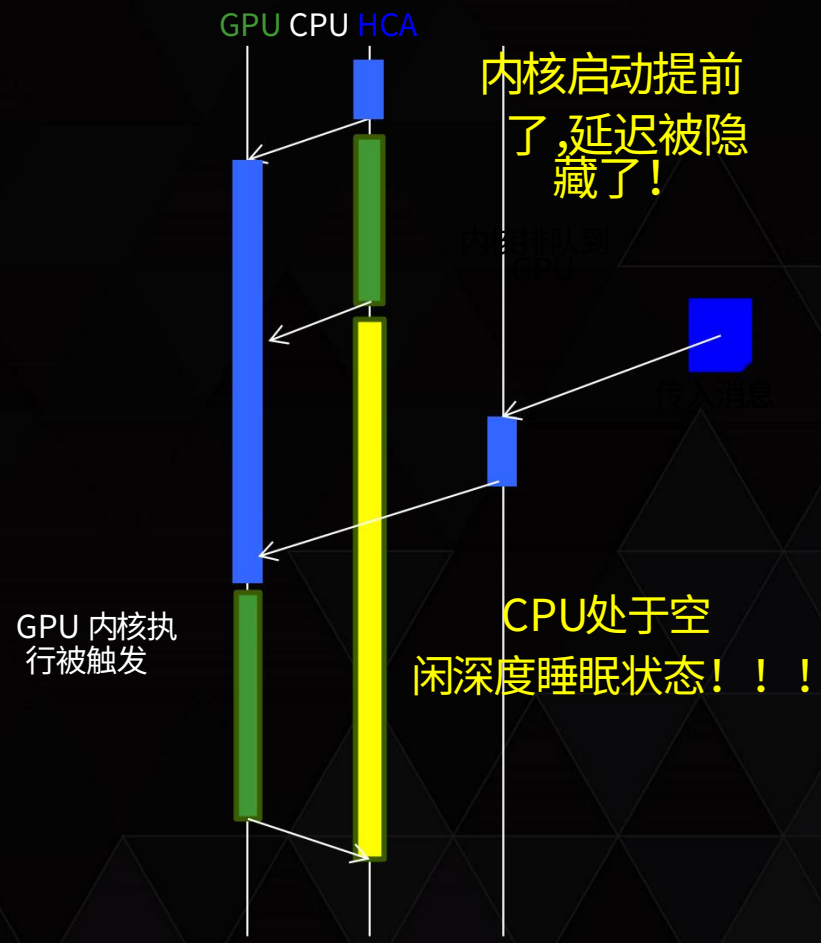
正常流量

```
while (!done) ibv_poll_cq();
a_kernel<<<...,stream>>>(buf);
cuStreamSynchronize(流);
```



接收+内核
GPU直接异步

```
gds_stream_wait_cq(流,rx_cq);
a_kernel<<<...,stream>>>(buf);
cuStreamSynchronize(流);
```



用例场景

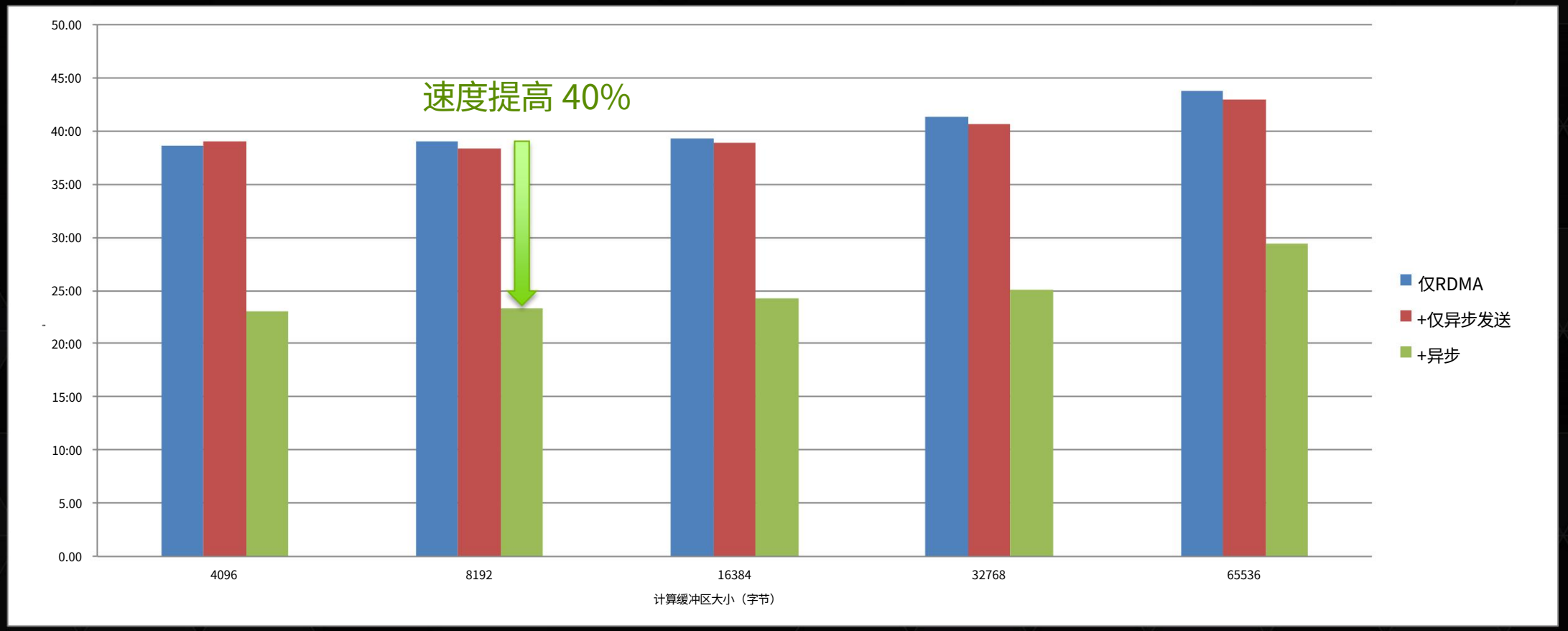
性能模式（~ Top500)启用批处理以

- 提高性能
-
- CPU 可用,额外 GFlops

经济模式（~ Green500）

- 启用 GPU IRQ 等待模式释放更多
- CPU 周期
- 可选更纤薄的CPU

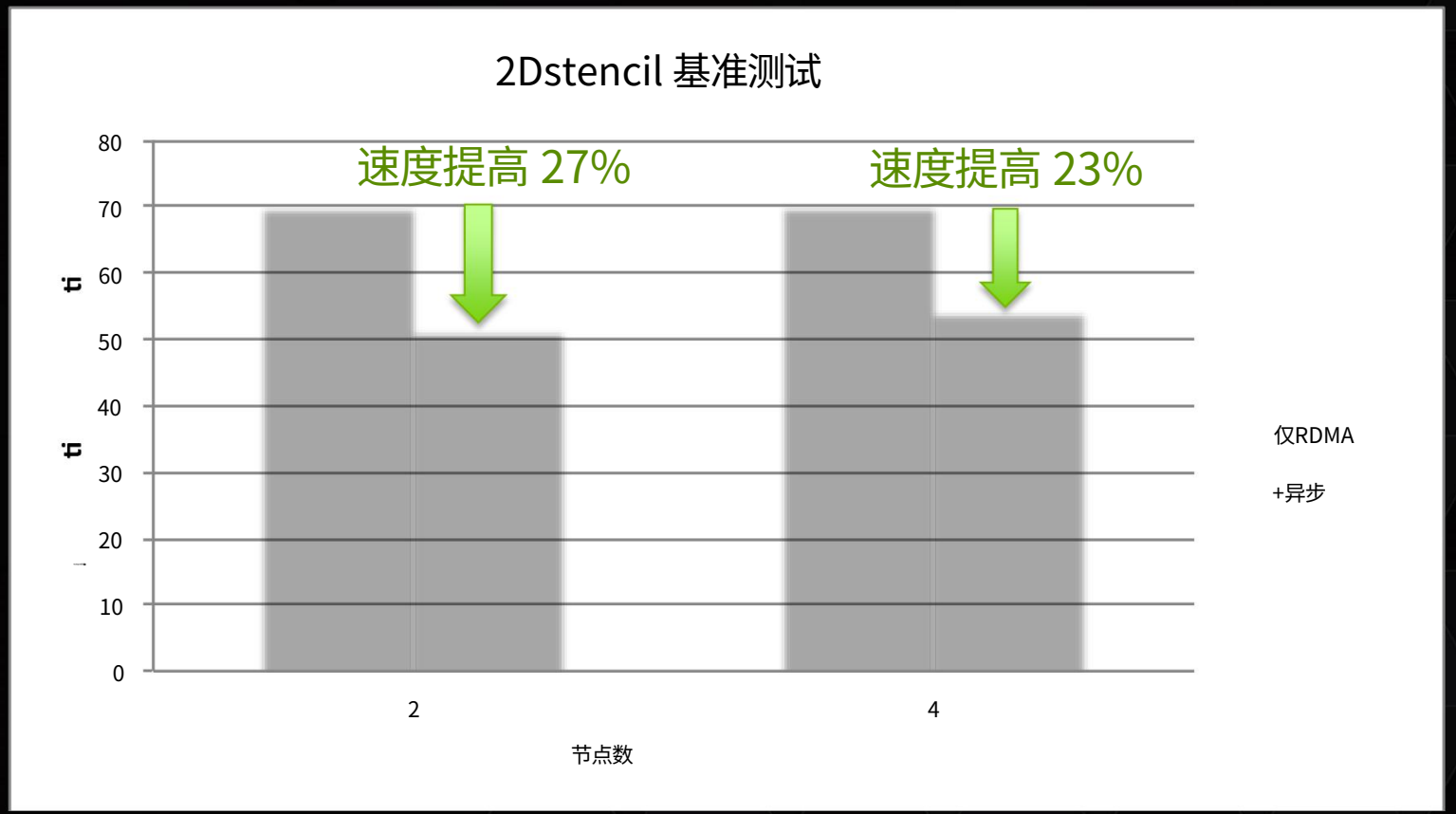
表演模式



[*] 修改了 ud_pingpong 测试:recv+GPU 内核+每一侧发送。 2个节点:Ivy Bridge Xeon + K40 + Connect-IB + MLNX交换机,10000次迭代,消息大小:128B,批量大小:20

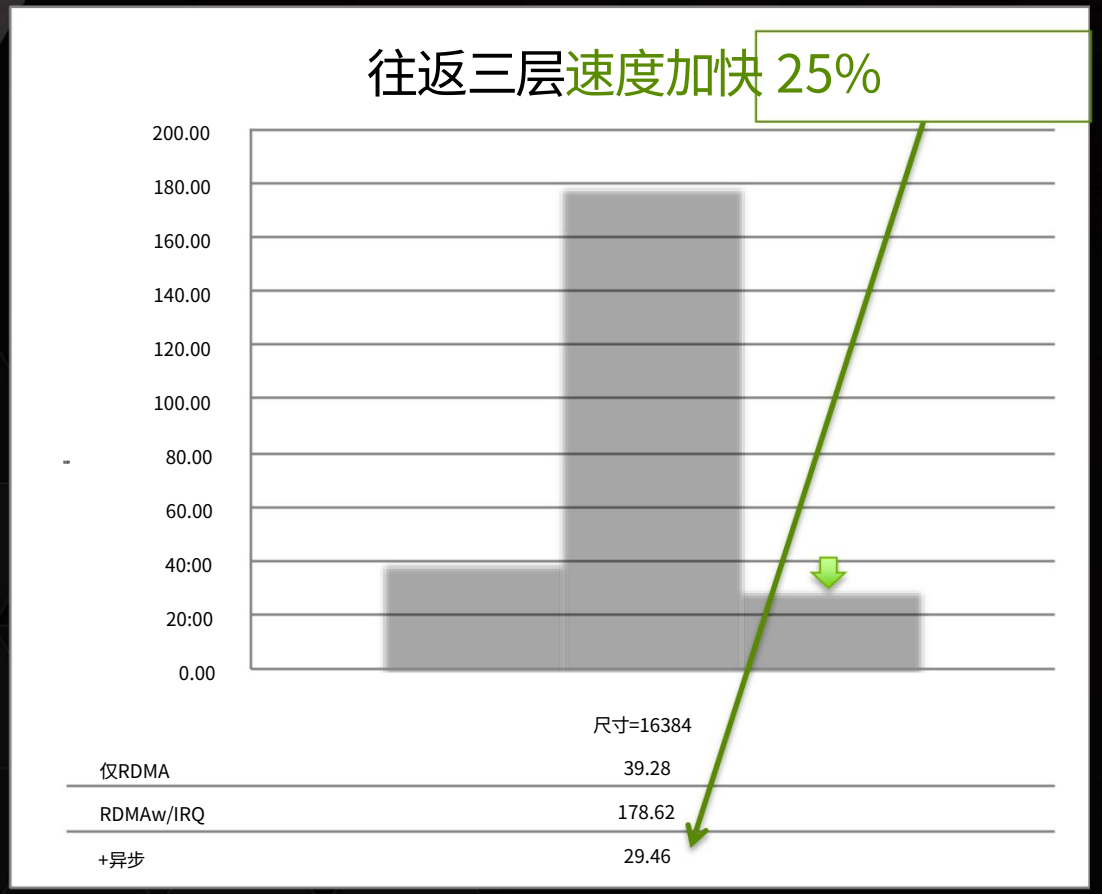
2D 模板基准

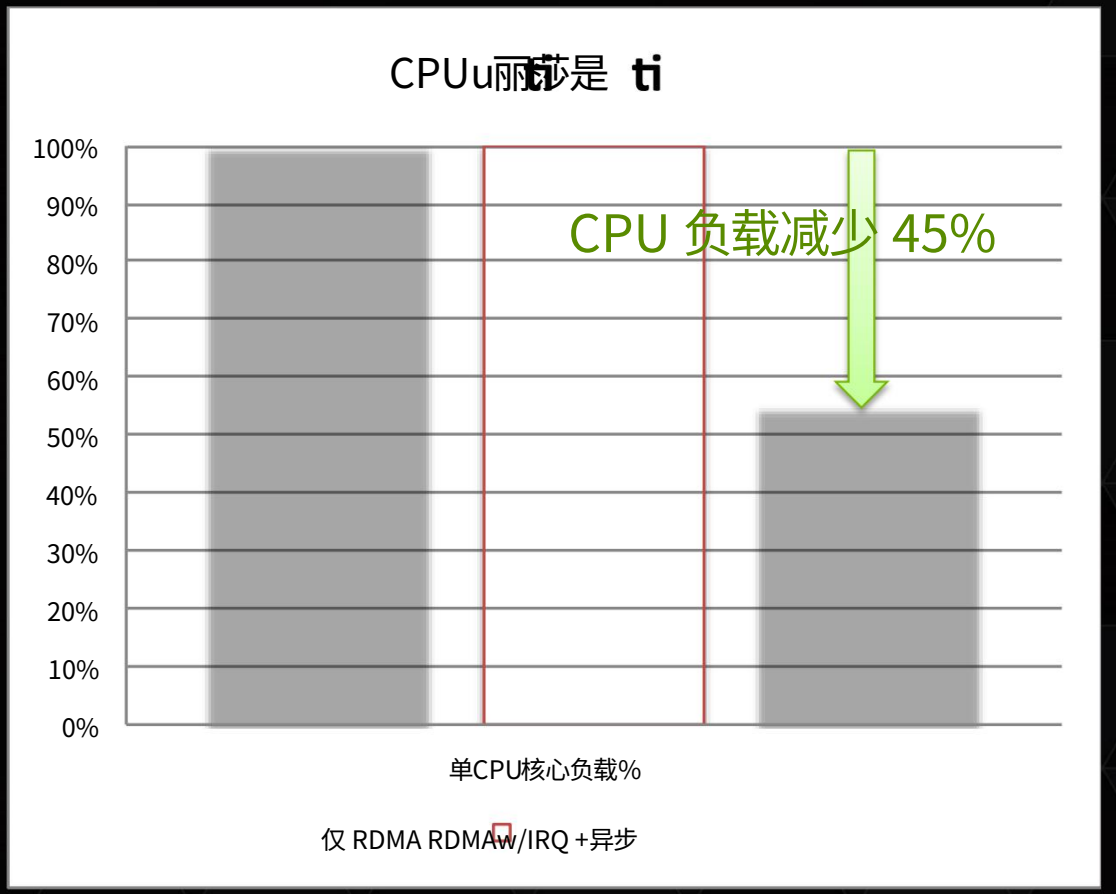
- 弱缩放
- 256^2 局部格子
- 2x1、2x2 节点网格
- 每个节点 1 个 GPU



[*] 4 个节点:Ivy Bridge Xeon + K40 + Connect-IB + MLNX 交换机

经济模式





[*] 修改了 ud_pingpong 测试,硬件与上一张幻灯片中的相同

概括

- 认识异步,下一代 GPUDirect
- GPU 协调网络操作
- CPU 脱离关键路径速度提
- 高 40%， CPU 负载减少 45%

对这些话题感到兴奋吗?合作与工作@NVIDIA

NVIDIA 注册开发者计划

- 使用 NVIDIA 产品进行开发所需的一切
- 成为会员是您与以下机构建立工作关系的第一步
  英伟达工程
  - 独家访问预发布版本
  - 提交错误和功能请求
  - 随时了解最新版本和培训机会
  - 获得独家下载
  - 独家活动和特别优惠
  - 在 NVIDIA 开发者论坛中与其他开发者互动

免费注册： developer.nvidia.com

GPU TECHNOLOGY CONFERENCE

谢谢

JOIN THE CONVERSATION

#GTC15

性能与经济性

性能模式

PowerTOP 2.3 Overview | Idle stats | Frequency

Package		CPU 0		
C0 polling	0.0%	C0 polling	0.0%	0.0 ms
C1-IVB	0.0%	C1-IVB	0.0%	0.0 ms
C3-IVB	0.0%	C3-IVB	0.0%	0.0 ms
C6-IVB	89.1%	C6-IVB	0.0%	0.0 ms
		CPU 1		
		C0 polling	0.0%	0.0 ms
		C1-IVB	0.0%	0.0 ms
		C3-IVB	0.0%	0.0 ms
		C6-IVB	98.8%	83.5 ms

经济模式

PowerTOP 2.3 Overview | Idle stats | Frequenc

Package		CPU 0		
C0 polling	0.0%	C0 polling	0.0%	0.0 ms
C1-IVB	0.8%	C1-IVB	7.9%	1.1 ms
C3-IVB	1.0%	C3-IVB	10.1%	1.1 ms
C6-IVB	91.3%	C6-IVB	23.2%	1.1 ms
		CPU 1		
		C0 polling	0.0%	0.0 ms
		C1-IVB	0.0%	0.0 ms
		C3-IVB	0.0%	0.0 ms
		C6-IVB	99.9%	126.1 ms

[*] 修改了 ud_pingpong 测试,硬件与上一张幻灯片相同,NUMA 绑定到 socket0/core0,SBIOS 省电配置文件