实验八：HPL
(High Performance Linpack)

郑海刚

HITSZ 实验与创新实践教育中心
Education Center of Experiments and Innovations, HITSZ

本讲概述

- CPU的浮点计算能力
- HPL介绍
- HPL.dat说明与调优
- 超算Top500介绍

CPU性能

- 同频性能
  - IPC：每时钟周期执行的指令数
  - CPI：每指令的执行时间时钟周期数
- 不同CPU的频率显然不同
  - MIPS（Million Instructions Per Second）：每秒执行的指令数
  - MIPS 衡量定点指令（整数操作）

计算机系统性能

- 除了CPU之外
  - 总线、内存、磁盘
  - SPEC测试
    - 图形渲染性能、邮件服务、存储、虚拟化等
- 由于应用的多样性，不同的计算机对不同的应用有不同的适应性，很难建立一个统一的标准来比较不同计算机的性能

FLOPS（floating point operations per second）

- 常用于科学计算领域，浮点操作比较多
- 每秒的浮点操作数，非指令数，不区分加减乘除
  - 一条指令可以有多个浮点操作，用指令数衡量不准确
  - 向量化指令（SIMD）一次可以执行多个数的操作，比如x86的avx256
  - SIMD的FMA（融合乘加）： c=a*b+c
  - CPU的理论性能如下

$$\mathbf{FLOPS} = \mathbf{cores} \times \frac{\mathbf{cycles}}{\mathbf{second}} \times \frac{\mathbf{FLOPs}}{\mathbf{cycle}}.$$

CPU基本信息查看与解读

- 查看CPU型号 lscpu 命令
  - 1 sockets : 1颗cpu
  - cores per socket：4核
  - threads per core：每核2个线程，采用了超线程技术，8个逻辑线程
  - 基准频率1.6G，实际运行频率可变
  - flags 包含了指令集的信息： avx等

```
$ ~ »lscpu
Architecture:                    x86_64
CPU op-mode(s):                  32-bit, 64-bit
Byte Order:                      Little Endian
Address sizes:                   39 bits physical, 48 bits virtual
CPU(s):                          8
On-line CPU(s) list:             0-7
Thread(s) per core:              2
Core(s) per socket:              4
Socket(s):                       1
Vendor ID:                       GenuineIntel
CPU family:                      6
Model:                           142
Model name:                      Intel(R) Core(TM) i5-10210U CPU @ 1.60GHz
Stepping:                        12
```

- wikichip 、Intel 、AMD官网也可查处理器信息

CPU浮点性能计算

- 现代X86处理器峰值浮点性能按照AVX指令计算
  - AVX2：寄存器是256位，一次可以进行256/64=4个浮点操作
  - 有FMA，一条指令最多2条浮点操作，乘以2
  - 如果包含两个执行单元（EUs），能同时执行两条AVX指令，继续乘以2
  - 再乘以频率（AVX工作时的频率）、乘以物理核心数
- https://en.wikichip.org/wiki/flops

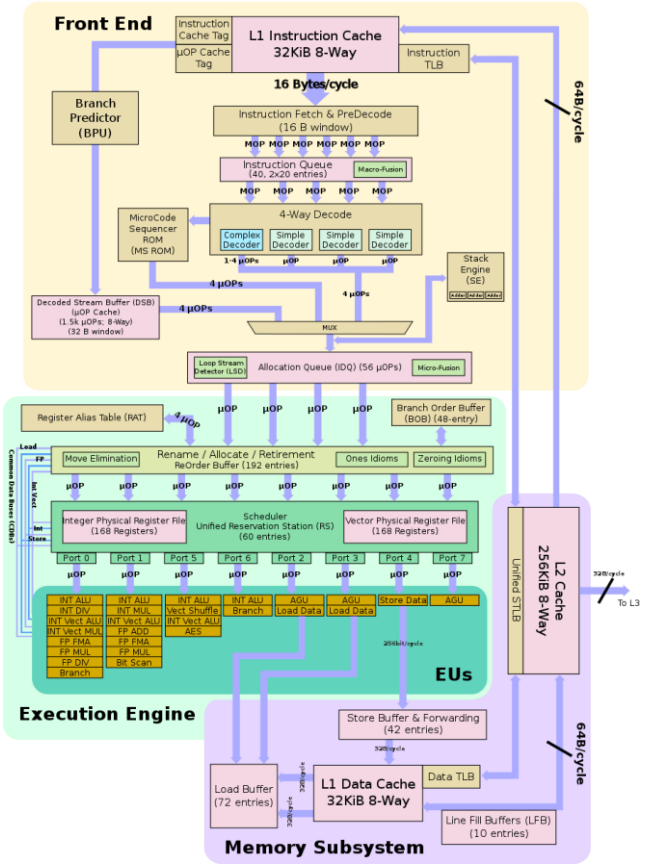
Haswell Broadwell Skylake Kaby Lake Amber Lake Coffee Lake Whiskey Lake	**EUs**	2 × 256-bit FMA		AVX2 & FMA (256-bit)
	DP	16 FLOPs/cycle	2 × 8 FLOPs	
	SP	32 FLOPs/cycle	2 × 16 FLOPs	

浮点计算单元数、处理器频率

- haswell 架构图
- 处理器的频率跟工作状态有关：
  - i5-4570R，单核频率3.2GHz
  - 4核全开3Ghz

List of Haswell Processors

				Turbo Boost				
Model	.2$	L3$	TDP	Frequency	1 Core	2 Cores	3 Cores	4 Cores
i5-4570R	MiB	4 MiB	65 W	2.7 GHz	3.2 GHz	3.1 GHz	3 GHz	3 GHz
i5-4670R	MiB	4 MiB	65 W	3 GHz	3.7 GHz	3.6 GHz	3.5 GHz	3.5 GHz
i7-4750HQ	MiB	6 MiB	47 W	2 GHz	3.2 GHz	3.1 GHz	3 GHz	3 GHz
i7-4760HQ	MiB	6 MiB	47 W	2.1 GHz	3.3 GHz	3.2 GHz	3.1 GHz	3.1 GHz
i7-4770HQ	MiB	6 MiB	47 W	2.2 GHz	3.4 GHz	3.3 GHz	3.2 GHz	3.2 GHz
i7-4770R	MiB	6 MiB	47 W	3.2 GHz	3.9 GHz	3.8 GHz	3.7 GHz	3.7 GHz
i7-4850EQ	MiB	6 MiB	47 W	1.6 GHz	3.2 GHz	3.1 GHz	3 GHz	3 GHz



服务器芯片

- 核心数更多，支持AVX512
- Intel Xeon Gold 6150

Mode	Base	Turbo Frequency/Active Cores										
		1	2	3	4	5	6	7	8	9	10	11
Normal	2,700 MHz	3,700 MHz	3,700 MHz	3,500 MHz	3,500 MHz	3,400 MHz	3,400 MHz	3,400 MHz	3,400 MHz	3,400 MHz	3,400 MHz	3,400 MHz
AVX2	2,300 MHz	3,600 MHz	3,600 MHz	3,400 MHz	3,400 MHz	3,300 MHz	3,300 MHz	3,300 MHz	3,300 MHz	3,300 MHz	3,300 MHz	3,300 MHz
AVX512	1,900 MHz	3,500 MHz	3,500 MHz	3,300 MHz	3,300 MHz	3,200 MHz	3,200 MHz	3,200 MHz	3,200 MHz	2,900 MHz	2,900 MHz	2,900 MHz

Linpack基准测试

- 通过求解线性方程组 $Ax=b$ 测量系统的浮点计算性能
- Linpack100：求解规模为100阶的稠密线性代数方程组
- Linpack1000：求解规模为1000阶的线性代数方程组，可以在不改变计算量的前提下做算法和代码的优化。
- HPLinpack：求解规模n可以自行调整，适用于并行系统

HPLinpack（ HPL测试）

- HPL（High-Performance Linpack Benchmark ）
  - 通过LU分解求解一个稠密线性方程组测试64位浮点峰值性能
    - 依赖BLAS库
  - 专注于分布式内存系统的性能测试
    - 依赖MPI实现
  - 可以调整矩阵规模，可改代码优化，但计算量要保证不变

HPL.dat参数说明

- **HPL Tuning**：上一行决定了下一行多少个数据有效
  - 1-2行：说明信息
  - 3-4行：内容输出到哪里
  - 5-6行：矩阵规模的大小
  - 7-8行：分块计算时块大小
  - 9行：处理器布局，推荐行优先
  - 10-12：进程网格的数量，乘积是总进程数
  - 14-21： 算法有关，LU分解的方式

```
1  HPLinpack benchmark input file
2  Innovative Computing Laboratory, University of Tennessee
3  HPL.out      output file name (if any)
4  6            device out (6=stdout,7=stderr,file)
5  1            # of problems sizes (N)
6  29 30 34 35  Ns
7  1            # of NBs
8  1 2 3 4      NBs
9  0            PMAP process mapping (0=Row-,1=Column-major)
10 2            # of process grids (P x Q)
11 2 1         Ps
12 2 1         Qs
13 16.0         threshold
14 3            # of panel fact
15 0 1 2        PFACTs (0=left, 1=Crout, 2=Right)
16 2            # of recursive stopping criterium
17 2 4          NBMINs (>= 1)
18 1            # of panels in recursion
19 2            NDIVs
20 3            # of recursive panel fact.
21 0 1 2        RFACTs (0=left, 1=Crout, 2=Right)
22 1            # of broadcast
23 0            BCASTs (0=1rg,1=1rM,2=2rg,3=2rM,4=Lng,5=LnM)
24 1            # of lookahead depth
25 0            DEPTHs (>=0)
26 2            SWAP (0=bin-exch,1=long,2=mix)
27 64           swapping threshold
28 0            L1 in (0=transposed,1=no-transposed) form
29 0            U  in (0=transposed,1=no-transposed) form
30 1            Equilibration (0=no,1=yes)
31 8            memory alignment in double (> 0)
```

HPL.dat参数常用调整策略

- FAQ： https://www.netlib.org/benchmark/hpl/faqs.html
- 自动调参工具
  - 节点数、CPU核数、内存、块大小

超算Top500

- 始于1993年，一年更新2次，每年6月的ISC大会，11月的SC大会
- 根据HPL测试数据排名
  - Rmax：HPL实测峰值，排名的依据
  - Rpeak：理论峰值

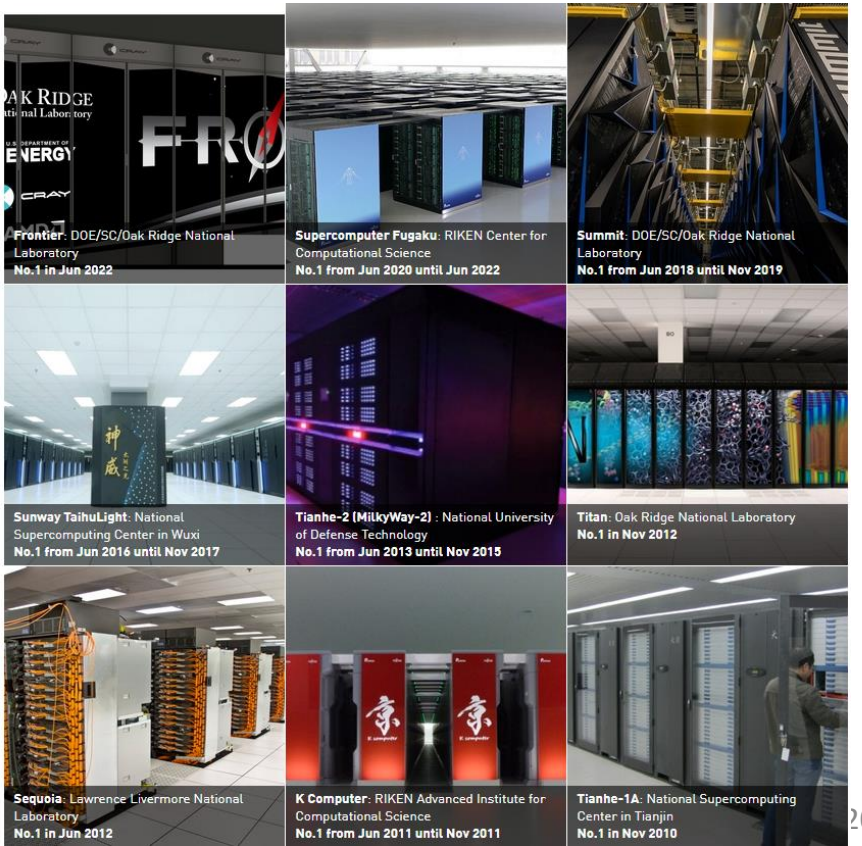
Top1：2024.6月

- Top1：美国Frontier
  - AMD EPYC 64C 2GHz
  - AMD Instinct MI250X
  - 8,699,904核数
  - Rmax：1206Pflops
  - Rpeak：1714Pflops

Rank	System	Cores	Rmax (PFlop/s)	Rpeak (PFlop/s)	Power (kW)
1	**Frontier** - HPE Cray EX235a, AMD Optimized 3rd Generation EPYC 64C 2GHz, AMD Instinct MI250X, Slingshot-11, HPE DOE/SC/Oak Ridge National Laboratory United States	8,699,904	1,206.00	1,714.81	22,786
2	**Aurora** - HPE Cray EX - Intel Exascale Compute Blade, Xeon CPU Max 9470 52C 2.4GHz, Intel Data Center GPU Max, Slingshot-11, Intel DOE/SC/Argonne National Laboratory United States	9,264,128	1,012.00	1,980.01	38,698
3	**Eagle** - Microsoft NDv5, Xeon Platinum 8480C 48C 2GHz, NVIDIA H100, NVIDIA Infiniband NDR, Microsoft Azure Microsoft Azure United States	2,073,600	561.20	846.84	
4	**Supercomputer Fugaku** - Supercomputer Fugaku, A64FX 48C 2.2GHz, Tofu interconnect D, Fujitsu RIKEN Center for Computational Science Japan	7,630,848	442.01	537.21	29,899
5	**LUMI** - HPE Cray EX235a, AMD Optimized 3rd Generation EPYC 64C 2GHz, AMD Instinct MI250X, Slingshot-11, HPE EuroHPC/CSC Finland	2,752,704	379.70	531.51	7,107

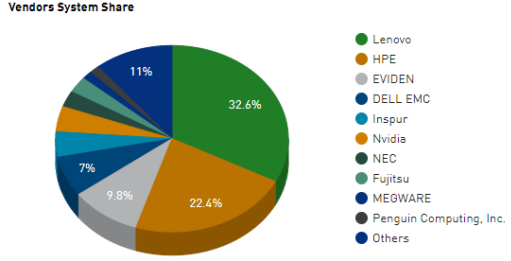
历史上的Top1

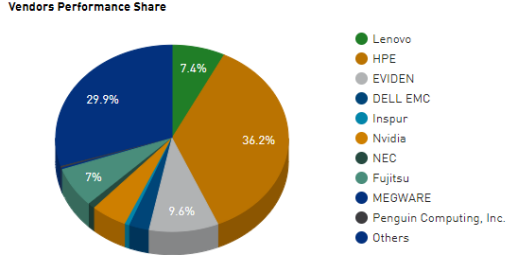
- 2010.11：天河1A
- 2013.6-2015.11：天河2
- 2016.6-2017.11：神威太湖之光



供应商份额

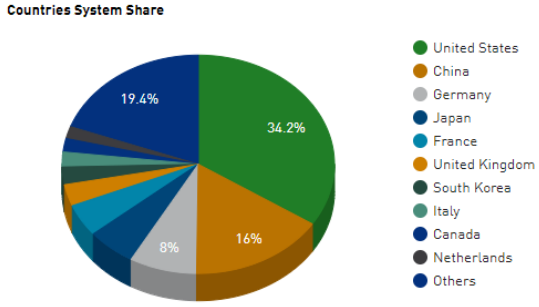
- 数量上：联想第一（32.6%），HPE第二（22.4%），浪潮第五（4.4%）
- 总算力占比：HPE第一（前身为惠普的企业级产品部门）
- 处理器Intel、IBM Power处理器为主
- 加速器Nvidia、AMD为主

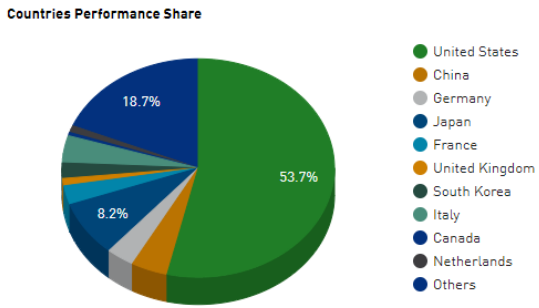




地区占比

- 数量上：美国（34.2%），中国（16%）
- 总的算力上：美国（53.7），中国（4.3%）





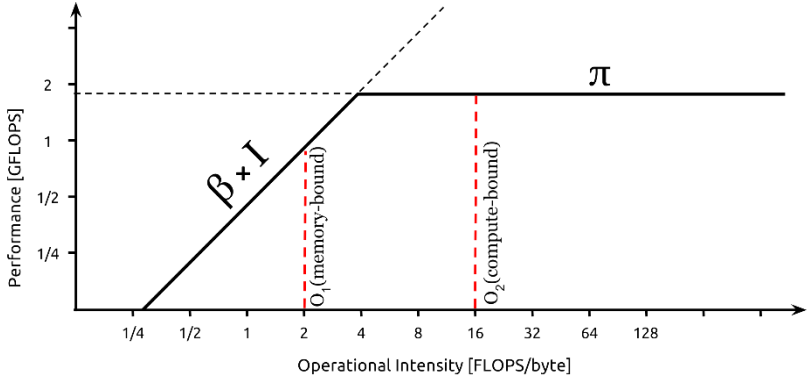
ASC超算竞赛HPL测试

- 在3kw功耗下，使用组委会提供的服务器，自带GPU卡，现场组装进行HPL测试
- 2023年取得了超过了100Tflops记录
- GPU也能跑HPL，用英伟达的容器方案

Roofline model（屋顶线模型）

- GEMM最快能有多快？系统的性能上限
- 估计应用在指定的硬件系统上能达到的性能、硬件的边界、潜在的优化
- 横轴是计算强度（Operational Itensity）：计算次数/访存字节数
- 纵轴是性能（ Performance ）

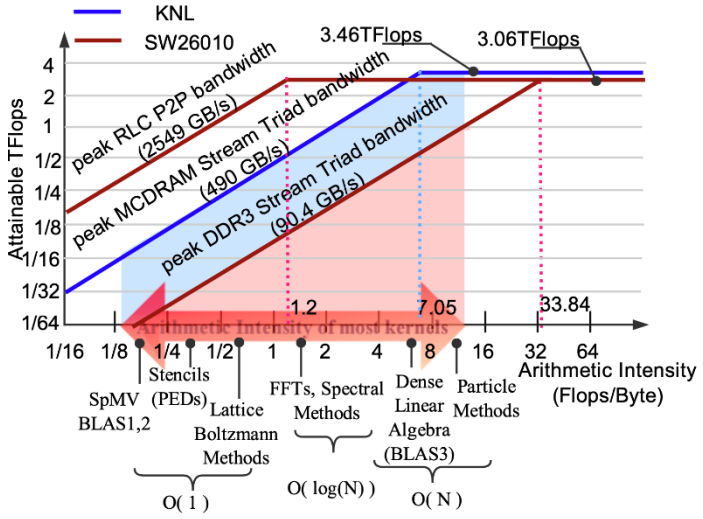
$$P = \min \begin{cases} \pi \\ \beta \times I \end{cases}$$



- *I*是计算强度，*π*是性能上限，*β*是内存带宽

内存带宽的2个例子

- 神威超算：内存带宽小，应用优化难



- GPU上的HBM（High Bandwidth Memory）

Type	Release	Clock (GHz)	Stack	per Stack (1024 bit)	
				Capacity (2^{30} Byte)	Data rate (GByte/s)
HBM 1	Oct 2013	0.5	8× 128 bit	1× 4 = 4	128
HBM 2	Jan 2016	1.0...1.2		1× 8 = 8	256...307
HBM 2E	Aug 2019	1.8		2× 8 = 16	461
HBM 3	Oct 2021	3.2	16× 64 bit	2×12 = 24	819
HBM 4	2026	5.6		2×16 = 32	1434

论文：Benchmarking SW26010 Many-Core Processor