链滴

# 架构师之路 - 服务器硬件扫盲

作者：jianzh5

原文链接：https://ld246.com/article/1599461115585

来源网站：链滴

很多架构师都是从软件开发成长起来的，大家在软件领域都有很深的造诣，大部分人对硬件接触的很。而成为架构师后需要频繁的跟人 、硬件 、软件 、网络打交道，本篇文章就给大家带来服务器硬件面的相关知识，主要包括服务器、CPU、内存、磁盘、网卡。

# 服务器

根据服务器的外形和使用场景我们将服务器分为以下四种：

- 塔式服务器

塔式服务器类似于台式机，主要适用于没有机房机架的公司，一般存放于中小办公环境。

- 机架式服务器

机架式服务器需要放置在标准机柜中，多存放于数据中心。

- 刀片服务器

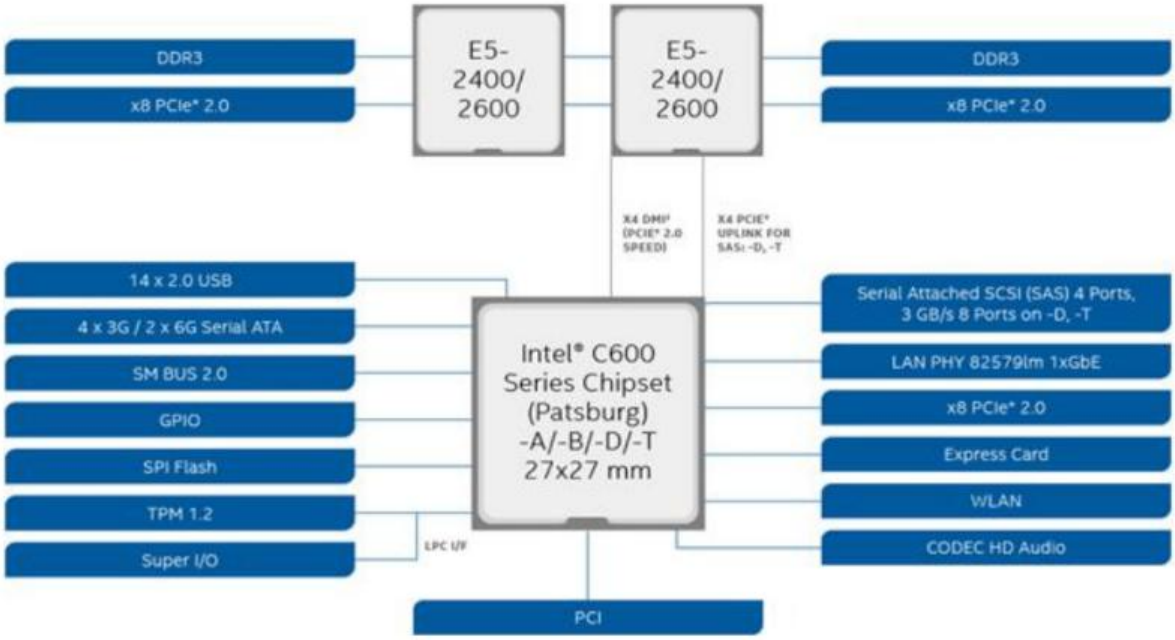
刀片服务器为了提供更高的密度，它比机架式服务器更节省空间，同时，散热问题也更突出，往往要机箱内装上大型强力风扇来散热，一般应用于大型的数据中心或者需要大规模计算的领域。

- 高密度服务器

高密度服务器是为了实现更高的空间利用率。

# 主板架构

服务器的机箱只是外壳，核心架构还是主板。




这个图是Intel典型的主板芯片组架构：服务器的主板有个统一的中央芯片组（Intel C600），芯片组以连接多个CPU（E5-2400/2600），CPU之间通过QPI快速通道进行连接，CPU与内存插槽，PCIE槽之间连接，芯片组还与低速的外设进行连接（USB、网卡、SATA等）。

## 服务器选型考虑因素

做服务选型主要基于以下几个方面：

- 限制条件：操作系统OS、客户喜好及预算、应用系统的编程语言
- 部署规模：如果规模较大，需要考虑空间占用问题，可考虑刀片或高密度服务器
- 扩展性：内存数量、磁盘数量、PCI插槽数量
- 稳定性
    - OS：UNIX > Linux > Windows
    - 硬件：小型机 > x86服务器
- 物理机、虚拟机、容器
    - 计算特点的考虑：是要将一个大的计算能力进行分割，灵活分配，还是有一个很大的课题要用台机器联合计算
    - IO特点的考虑：吞吐率与IOPS多大？虚拟机能否承受？

## 服务器厂商

- 国内的服务器厂商主要有：曙光、华为、浪潮、H3C、联想、长城等
- 国外的服务器厂商主要有：Dell、HP、IBM

# CPU

CPU作为服务器的核心固件，我们主要通过以下几个概念来了解：

- Socket

Socket俗称多少路，就是一个服务器主板上可以安装几个物理CPU

- Core

一个物理CPU实际可以有几个内核（Core），比如我们经常听到的32核64核、128核等等

- 超线程

如果一个内核可以同时运行2个线程我们就称这个CPU具有超线程能力，反之则不具备超线程能力

- 频率

也叫主频，这个越高越好

- 内存通道

每个CPU能支撑的最大内存数，Intel最新的能支持6个

- 内存带宽、内存规格

CPU支持什么规格的内存，支持的频率范围是多少

以上的这些概念数据可以通过CPU的产品规格书中进行详细了解。在服务器上可以通过 lscpu 命令查cpu信息

```
[root@oaup prop]# lscpu
Architecture:          x86_64
CPU op-mode(s):        32-bit, 64-bit
Byte Order:            Little Endian
CPU(s):                72
On-line CPU(s) list:   0-71
Thread(s) per core:    2
Core(s) per socket:    18
Socket(s):             2
NUMA node(s):          2
Vendor ID:             GenuineIntel
CPU family:            6
Model:                 85
Model name:            Intel(R) Xeon(R) Gold 6240 CPU @ 2.60GHz
Stepping:              7
CPU MHz:               1000.000
BogoMIPS:              5198.44
Virtualization:        VT-x
L1d cache:             32K
L1i cache:             32K
L2 cache:              1024K
L3 cache:              25344K
NUMA node0 CPU(s):     0-17,36-53
NUMA node1 CPU(s):     18-35,54-71
```

上图服务器CPU显示有72个，是因为有2个Socket，每个Socket有18核而每核可以同时运行2个线程通过 2 x 18 x 2 = 72 得到。

# 厂商

- 国外的CPU厂商主要有：Intel、AMD
- 国内的CPU厂商主要有：龙芯、兆芯、飞腾、海光、申威、华为等

# 主流产品介绍

## Intel系列

Intel现在主推的是“Intel至强可扩展”系列，在这个系列下又分为四档：铂金、金、银、铜，每档下又有不同的型号


## AMD系列

AMD系列主要用到的是霄龙系列，霄龙系列CPU核数很高，下面我们看看几款具体的产品

| 型号 | CPU核心数 | 线程数量 | 最大加速时钟频率 | 准时钟频率 | 默认TDP/TDP |
|---|---|---|---|---|---|
| 霄龙7742 | 64 | 128 | 高达3.4GHZ | 2.5GHZ | 225W |
| 霄龙7702 | 64 | 128 | 高达3.35GHZ | 2HZ | 200W |
| 霄龙7702P | 64 | 128 | 高达3.4GHZ | GHZ | 200W |
| 霄龙7642 | 48 | 96 | 高达3.35GHZ | 2.GHZ | 225W |
| 霄龙7552 | 48 | 96 | 高达3.3GHZ | 2.2HZ | 200W |
| 霄龙7542 | 32 | 64 | 高达3.4GHZ | 2.9HZ | 225W |

## 华为鲲鹏系列

### 鲲鹏916（低功耗级）

- 32核/2.4GHz/16nm/75W
- 4通道DDR4控制器
- PCIe 3.0，10GE
- 支持2路互联

### 鲲鹏920-3326/4826（极致效能型）

- 32/48核/2.6GHz/7nm/120/150W
- 8通道DDR4控制器
- PCIe 4.0，100GE,CCIX
- 支持2/4路互联

**鲲鹏920-3326/4826（极致性能级）**

- 64核/2.6GHz/7nm/180W
- 8通道DDR4控制器
- PCIe 4.0 ， 100GE,CCIX
- 支持2/4路互联

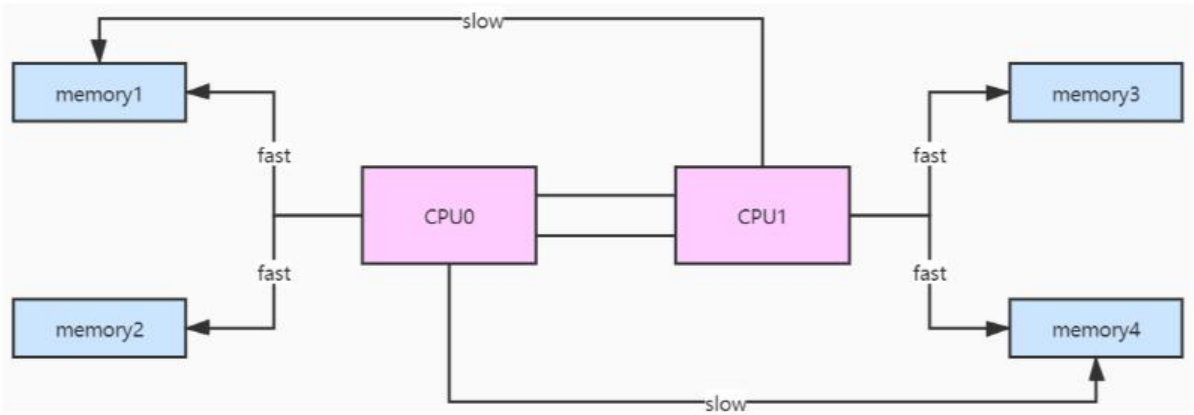
# 主流CPU型号性能横向对比

| CPU类型 | 笔记本CPU | 台式机CPU | 台式机CPU | 台式机CPU | 台式机CPU | 服务器CPU | 服务器CPU | 服务器CPU |
|---|---|---|---|---|---|---|---|---|
| CPU厂商 | INTEL | AMD | 龙芯 | 兆芯 | 飞腾 | INTEL | 飞腾 | 华为 |
| CPU型号 | i5-7200U | A10-7850K | 3A3000 | C4701 | FT1500A/4 | Gold 5118 | FT2000PLUS | Hi1616 |
| Socket | 1 | 1 | 1 | 1 | 1 | 2 | 1 | 2 |
| 逻辑CPU | 4 | 4 | 4 | 4 | 4 | 48 | 64 | 64 |
| SPECjvm2008 compress得分 | 118 | 177 | 71 | 74 | 81 | 2790 | 1343 | 1495 |

通过上图大家可以看到国产CPU与国外CPU之间的差距，性能基本只有Intel中档CPU性能的一半左右国产CPU还有很长的路要走。

# NUMA

NUMA 即 Non-Uiform Memory Access（非一致性内存访问），结合我们之前讲述的主板架构，颗CPU之间有一个通道，每个CPU与各自的内存通道进行直连，可以通过下图直观看出。




CPU0 访问 左边的内存通道速度很快，CPU1访问右边的内存通道也很快，当CPU1要访问左边的内通道必须要借助CPU0的帮忙，需要先通过QPI总线找到CPU0，再来访问左边的内存通道，这就产生额外的开销，访问左边内存通道的开销相当于直连访问右边通道开销的3倍。

**所以对于计算密集型任务我们需要尽量避免这种跨CPU的内存访问，这就是NUMA的问题，非一致指的是访问本地和跨CPU访问的代价差别不一致**

我们可以通过指令 numactl -s 查看numa的信息

```
[root@l72 ~]# numactl -s
policy: default
preferred node: current
physcpubind: 0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15
cpubind: 0 1
nodebind: 0 1
membind: 0 1
```

可以使用指令 numactl --cpunodebind=0 --membind=0 command 指定进程使用的numa节点和存

如上就是让 command 指令只使用cpu0 ，和内存0，这就使得进程指令在运行的时候使用的CPU和存在同一侧，达到计算性能速度最大化的效果。

# 内存

内存大家平时工作中接触的都比较多，对于内存我们主要通过以下几个方面来了解：

- 内存规格

DDR3、DDR4，目前主流已经是DDR4

- 内存大小

2G、4G、8G、16G、32G

- 内存频率

1333MHz, 1600MHz, 1866MHz、2133MHz, 2400MHz , 2666MHz

- 带宽

即CPU对内存实际读写数据的速度，DDR4 2400内存的带宽为30GB/s左右

- 通道

一个CPU可以连接多个内存，CPU上的内存通道数指的是CPU能并发访问直连多少个内存。4通道表示PU可以同时访问与之直连的4根内存，这样就能达到带宽翻四倍的效果。

在4通道模式下读取1G的数据进内存，最终数据会分布在4根内存上而不是一根内存，这就实现了速的翻4倍；

每颗CPU对自己的内存控制器直连的内存访问速度较快，要访问另一颗CPU连接的内存时，需要通过PI总线，开销为本地内存的3倍。

## 了解内存信息

主要通过以下三个命令全面了解内存信息

- 我们可以通过 dmidecode -t memory | more 指令查看内存信息，效果如下：

```
[root@172 ~]# dmidecode -t memory | more
# dmidecode 3.0
Getting SMBIOS data from sysfs.
SMBIOS 2.6 present.

Handle 0x1000, DMI type 16, 15 bytes
Physical Memory Array
        Location: System Board Or Motherboard
        Use: System Memory
        Error Correction Type: Multi-bit ECC
        Maximum Capacity: 288 GB
        Error Information Handle: Not Provided
        Number Of Devices: 18

Handle 0x1100, DMI type 17, 28 bytes
Memory Device
        Array Handle: 0x1000
        Error Information Handle: Not Provided
        Total Width: 72 bits
        Data Width: 64 bits
        Size: 8192 MB
        Form Factor: DIMM
        Set: 1
        Locator: DIMM_A1
        Bank Locator: Not Specified
        Type: DDR3
        Type Detail: Synchronous Registered (Buffered)
        Speed: 1333 MHz
        Manufacturer: 002C00B3802C
        Serial Number: DB643DBE
        Asset Tag: 08105161
        Part Number: 36JSZF1G72PZ-1G4D1
        Rank: 2
```

- 可以使用 dmidecode -t memory | grep Size 指令查看内存大小并判断内存是否正常工作

```
[root@172 ~]# dmidecode -t memory | grep Size
        Size: 8192 MB
        Size: 8192 MB
        Size: 8192 MB
        Size: No Module Installed
        Size: No Module Installed
        Size: No Module Installed
        Size: No Module Installed
        Size: No Module Installed
        Size: No Module Installed
        Size: 8192 MB
        Size: No Module Installed
        Size: 8192 MB
        Size: No Module Installed
        Size: No Module Installed
        Size: No Module Installed
        Size: No Module Installed
        Size: No Module Installed
        Size: No Module Installed
```

将内存插入主板时一般需要对称插入，通过上图我们可以看到下面有根内存不工作。

- 可以通过 free 指令查看内存容量

```
[root@172 ~]# free
              total        used        free      shared  buff/cache   available
Mem:       24507904      292424    23281040        8844      934440    23801652
Swap:      33566716           0    33566716
```

系统剩余内存 available 是我们最关心的一个值，不要被free列唬住了。

# 磁盘

对于磁盘我们主要通过吞吐率和IOPS两个指标来对其衡量

**吞吐率/吞吐量**：单位时间内读写的数据量

- 机械硬盘：约100MB/s – 200MB/s；
- 普通固态硬盘：200MB/s - 500MB/s；
- PCIE固态硬盘（直连CPU）：900MB/s - 3GB/s

**IOPS**：每秒IO操作的次数

- 机械硬盘：100-200
- 普通固态硬盘：30000-50000
- PCIE固态硬盘（直连CPU）：数十万

为什么很多性能比较慢的服务在软件层面进行优化收益很小，而更换一块固态硬盘就能解决所有问题问题就在这里。

普通固态硬盘的吞吐率大概为机械硬盘的23倍，而IOPS确达到了机械硬盘的25000倍。

**IOPS和数据吞吐量适用于不同的场合：**

在随机读写频繁的应用中，如OLTP(Online Transaction Processing)，IOPS是关键衡量指标。

对于大量顺序读写的应用，则更关注吞吐量指标。

读取10000个1KB文件，用时10秒 Throught(吞吐量)=1MB/s ， IOPS=1000 追求IOPS

读取1个10MB文件，用时0.2秒 Throught(吞吐量)=50MB/s, IOPS=5 追求吞吐量

# 网卡

网卡，又称网络适配器或网络接口卡，英文名为Network Interface Card。在网络中，如果有一台计机没有网卡，那么这台计算机将不能和其他计算机通信，它将得不到服务器所提供的任何服务了。当如果没有网卡，就称不上服务器了，所以说网卡是服务器必备的设备，就像普通PC（个人电脑）要配理器一样。

我们也可以也通过以下几个维度来了解下网卡：

**网卡速度规格**

100M、1G、10G、25G

**网卡接口类型**

RJ45（电、短距离）、光纤（光、长距离）

**网卡绑定模式**

多网卡绑定一方面能够提高网络吞吐量，另一方面也可以增强网络高可用。

从软件的角度来看，多网卡绑定实际上只需要提供一个额外的bond驱动程序即可，通过该虚拟网卡动程序可以将实际多块网卡屏蔽，对TCP/IP协议层而言只存在一个Bond网卡。Linux主要有7种绑定式：

- broadcast (广播策略：data is transmitted over all ports)

这种模式的特点是一个报文会复制两份往bond下的两个接口分别发送出去。当有对端交换机失效，们感觉不到任何丢包。

● round-robin (轮询策略：data is transmitted over all ports in turn)

该模式下，链路处于负载均衡状态，数据以轮询方式向每条链路发送报文，基于per packet方式发送即每条链路各一个数据包，这模式好处在于增加了带宽，同时支持容错能力，当有链路出问题，会把量切换到正常的链路上。

● active-backup (主备策略：one port or link is used while others are kept as a backup)

在该模式下，一个端口处于主状态，一个处于备状态，所有流量都在主链路上发出和接收，备链路不有任何流量。当主端口down掉时，备端口接管主状态。

● loadbalance (适配器传输负载均衡：with active Tx load balancing and BPF-based Tx portselectors)

在该模式下，通过源和目标mac做hash因子来做xor算法来选择链路，这样就使得到达特定对端的流总是从同一个接口上发出。

● lacp (动态链路聚合：implements the 802.3ad Link Aggregation Control Protocol)

在该模式下，操作系统和交换机都会创建一个聚合组，在同一聚合组下的网口共享同样的速率和双工定。

# 小结

本文给大家介绍了服务器硬件的基础知识，只有对硬件有了全面的认识和了解我们才能在硬件选型时到心中有沟壑，可以针对各个组件的特点选取合适的硬件来支撑其运行。