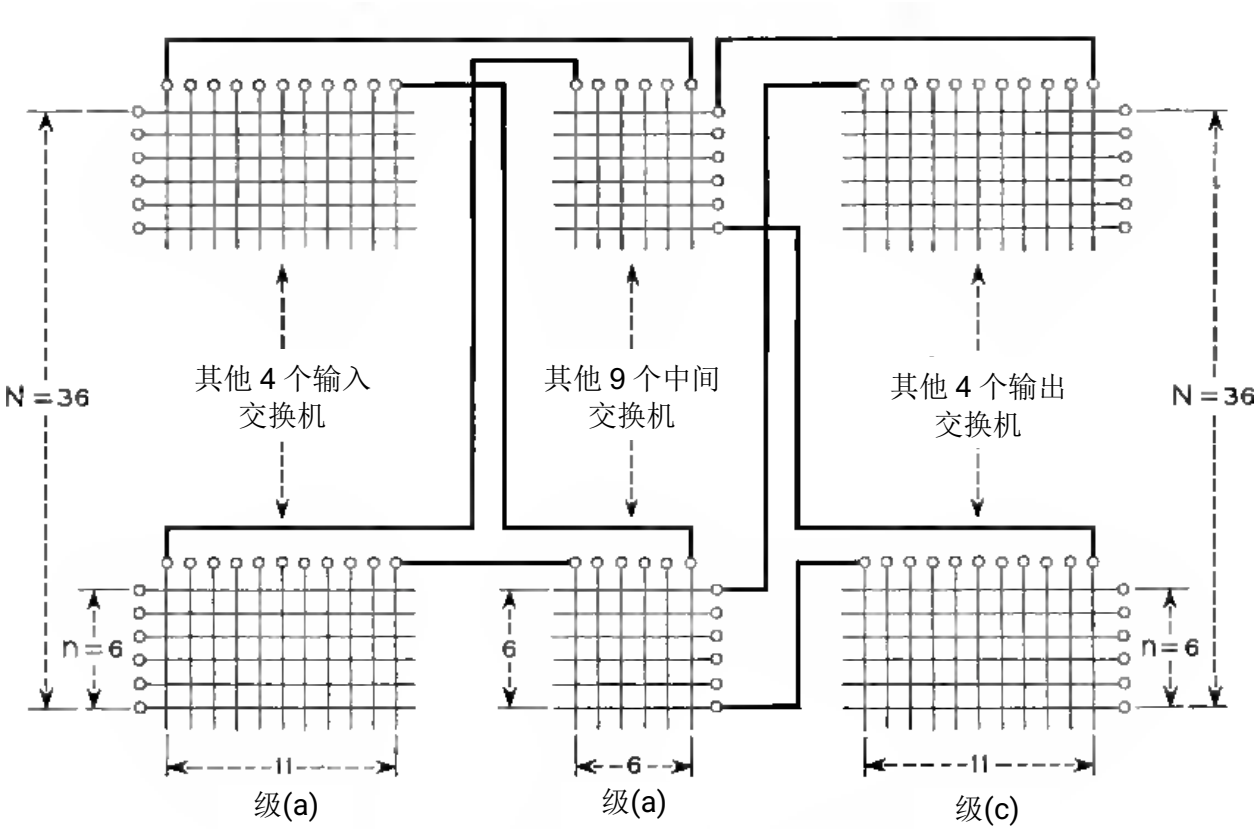
Clos 和机箱在运行人工智能应用方面真的有可比性吗？

为人所熟知且经常被提及的Clos拓扑可以追溯到70年前，一位名叫Charles
Clos的美国工程师在贝尔实验室工作时发表的一篇文章，文章名为“无阻塞交换网络的研究”。

这些选择都说不上简单或者理想。事实上，添加路由器的成本很高，并且会增加路由和中继的复杂性。简单地说，Clos拓扑是由具有预定义外部端口和内部端口的设备组成的纵横式交换矩阵（crossbar）。事实上，Clos拓扑有一种相当常见的实现方式，可以在机箱中看到。

抛开准确性不谈，Clos和机箱这两个选项在连接人工智能集群方面具有明显的差异。这篇博文将对这两种架构在运行人工智能应用方面进行的比较。



交叉点数量=6N3/2－3N（N=36 时交叉点数量为 1188）

图1 ：Clos拓扑

（来源：Charles Clos《无阻塞交换网络的研究》）

什么是Clos？

正如开篇所提到的那样，Clos是一种纵横式交换矩阵，具有进入这个结构的入口点、交叉点以及通过这个结构后的出口点。

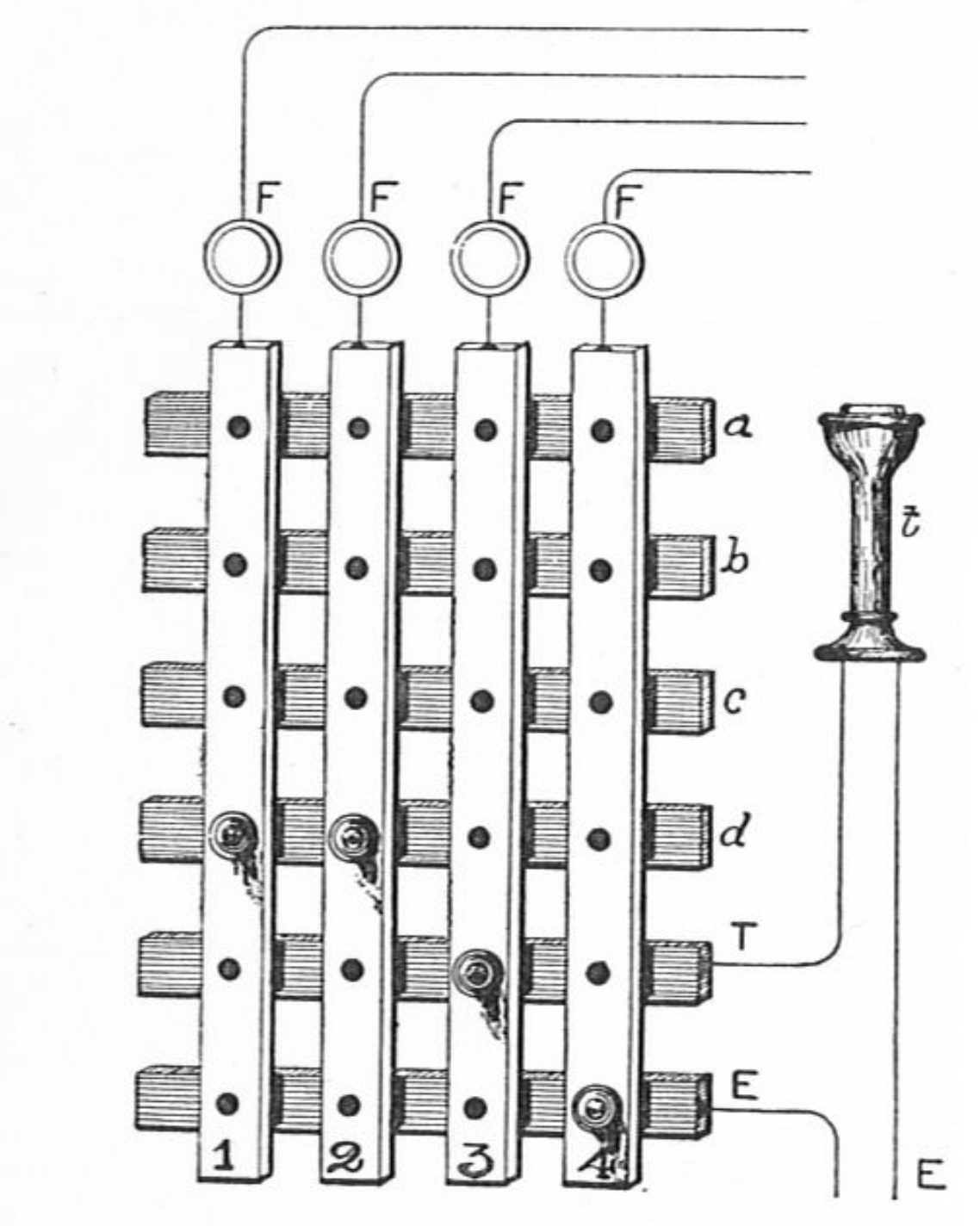


图2：纵横式交换矩阵电话交换机

引脚标记入口和出口之间的连接

（来源：维基百科）

在人工智能领域，Clos旨在连接为运行人工智能应用而构建的数据中心中的端点设备，其中这些端点指的是服务器，具体是安装在服务器上的网络接口卡（NIC）。Clos纵横式交换矩阵的入口点是这些服务器物理连接的交换机，通常位于安装服务器的机架顶部，令人惊讶的是它被称为架顶式（ToR）交换机。然后，这个ToR被连接到多个聚合交换机（称为脊交换机或结构或第2层交换机）。需要注意的是，这些脊交换机仅连接到ToR交换机以及网络中的所有ToR交换机。

进入Clos纵横式交换矩阵的流量在到达目标服务器之前将穿过ToR交换机、一个脊交换机和另一个ToR交换机。这种拓扑被视为是无阻塞的，因为纵横式交换矩阵的交叉对分带宽（所有脊交换机的聚合带宽）等于面向ToR端口的所有端点的聚合带宽。

理论上，如果所有端点在一段时间内传输全带宽（简单起见，我们假设允许这种情况的流量模式），则纵横式交换矩阵具有足够的容量让所有端点同时接收全带宽。这里的假设是，为每个流量选择的脊交换机是“正确的”脊交换机，因此当纵横式交换矩阵完全加载时，所有脊交换机都会处于100%利用率。不幸的是，这是不可能的……哪怕是理论上。在构建用于数据中心连接的Clos拓扑时，整个数据中心绝不会以“任意到任意流量模式”运行最大带宽。

我在上面的例子中没有提到的一点是指以太网技术。在以太网中，入口ToR需要根据要使用的脊交换机来决定每个流量。这种流量管理基于哈希算法，这类算法应该尽可能接近随机。

另一种能够实现相同拓扑的技术是InfiniBand。在这种情况下，ToR决定的脊交换机选择是基于外部“大脑”提供的附加信息，“大脑”对整个纵横式交换矩阵具有完全可见性。这让充分利用纵横式交换矩阵的带宽变得更容易实现，但还需要有关运行于整个服务器集群中的流量类型的专业知识。可以将之想象为带着提

供方向的空中悬浮摄像头穿越迷宫。虽然它可以带你穿过迷宫，但你需要一个悬空的摄像头提供方向……

因此，当被问到你是否使用Clos或InfiniBand时，答案应该是“是”。

当被问到你是否使用Clos或支持智能NIC的网络时，答案应该是“是”。

当被问到你是否使用Clos或机箱时，答案应该是“是”。

机箱是如何打造的？

清单如下：

- 一个路由引擎（通常是两个，另一个用作冗余），用于构成系统的大脑
- 几个电源和风扇
- 连接到端点和脊交换机卡的线路卡（在机箱可容纳范围内越多越好）
- 仅连接到线路卡的脊交换机卡（或结构卡）
- 将所有线路卡连接到所有脊交换机卡的物理纵横式交换矩阵
- 用于容纳所有这些组件的金属外壳

简而言之，“机箱”就是“装进盒子里的Clos”。

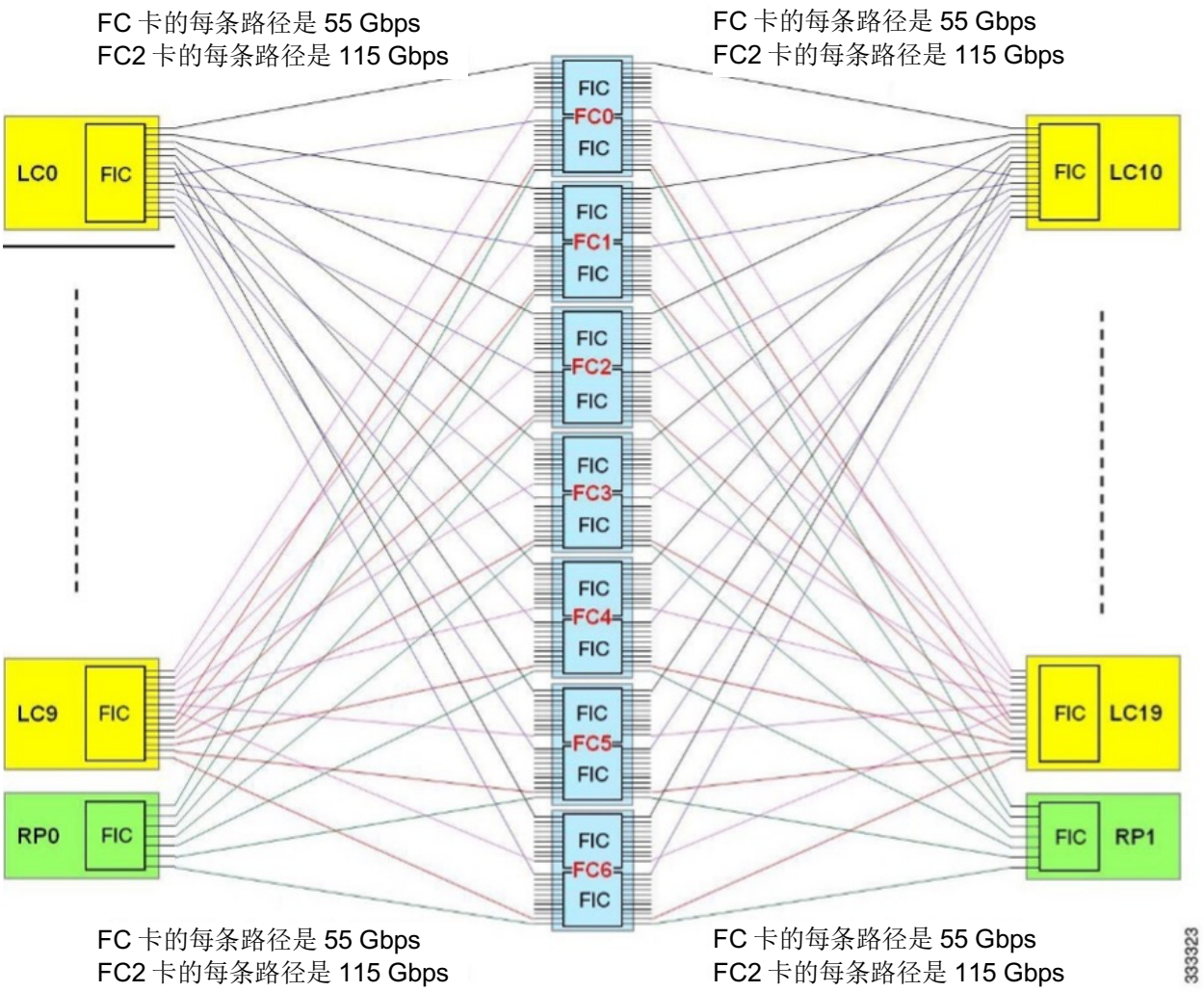


图 3：内部机箱拓扑
（来源：53bits）

组网拓扑不仅出色，还易于操作

我们已经确定，我们并不是在比较Clos和类似的机箱拓扑。现在让我们看看真正让两种模型成为优先选择的属性.....先从机箱开始。

机箱易于操作，在最佳工作点运行时非常高效。单一控制平面。单一IP地址。用户可以选择适合自己部署的尺寸，然后只需以热插入模式装上线路卡。出现

问题时，只需向一个供应商问责。在所有线路卡和端口上全速运行时，其每千兆流量的功耗最低。

唯一的问题是永远无法达到这个最佳工作点。当机箱利用率达到约70%时，就会进行更换，因为流量高峰可能会使利用率超过100%。机箱的固定性决定了在扩展方面它无法超出自己的金属外壳。用户可以从尽可能大尺寸的机箱开始，但这意味着大多数时候机箱的利用率很低，而达到较高利用率时，看到的将会是遍布整个数据中心的杂乱布线，因为所有这些布线都连接到这一台设备。

不过，由于内置的内部机制，机箱能够提供最可预测的行为，而这些机制是无法在通用网络中实现的。这对于人工智能工作负载至关重要。

Clos（这里指的是著名的以太网Clos）表现出色之处，正是机箱所不能及之处。它由小型交换机组成，通常成本较低。它非常易于在网络中进行扩展和改动。它是横向而不是纵向扩展。在构建网络时，用户可以与多个供应商合作，如此可以更好地控制供应链，最终控制价格。

然而，性能仍有一些不足之处。机箱会吸收流量并将其分段，以“喷射”到所有结构元件，然后在接收端重组；Clos中的流量处理远不如机箱内置的内部机制那么高效。

此外，Clos中的每个元件同时也是网络中的一个元件。因此，在大型部署中，用户最终可能需要管理数以百计的设备，而且它们之间的网络也需要管理、监控、保护、故障排除和修复。所有这些都必须尽快完成，

而在被迫需要选择时，总会有所妥协。服务提供商更青睐机箱的性能，而数据中心则渴望Clos的规模。

在数据中心运行的人工智能工作负载则两者都需要。

什么是Distributed Disaggregated Chassis (DDC) 操作模型

Distributed Disaggregated Chassis (DDC)
实际上是一种没有金属外壳的机箱。

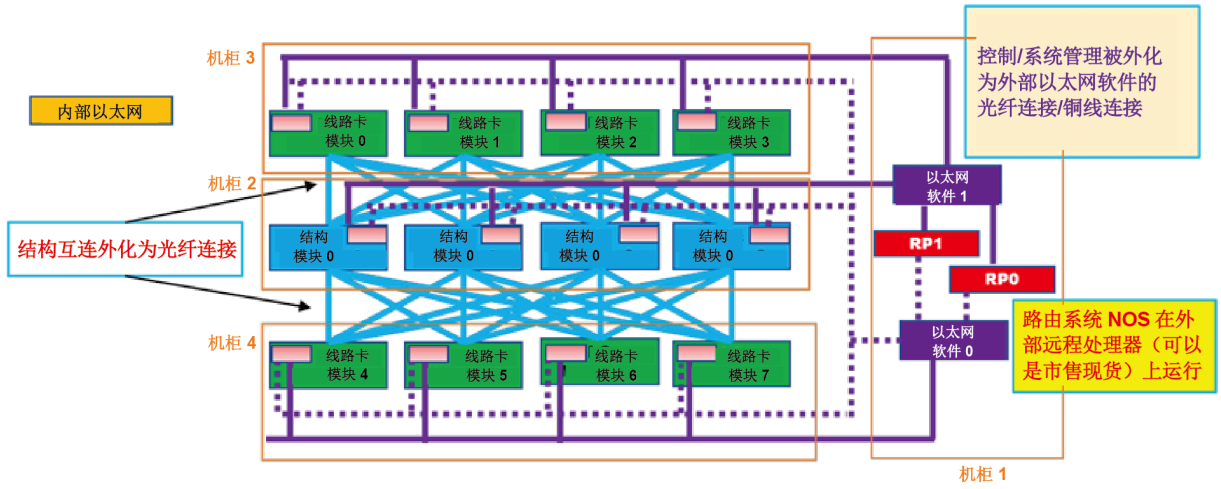


图4：DDC架构
（来源：DDC的OCP版本1规格）

DDC的构建采用了与机箱相同的组件。在这种情况下，它不是内置在单个容器中，而是分布在多个独立设备中，这些设备充当的便是“线路卡”和“脊交换机（结构）卡”。这种解耦化（disaggregation）属性使这些独立设备能够作为标准白盒，与运行它们的软件分开采购。

目前看来，这似乎是典型的Clos，不过DDC多了另一个元件，也即系统的大脑（增加一个用作冗余），它对白盒进行控制，同时保存和共享网络的相互“知识”。这意味着DDC在两个关键方面表现为一个统一的元件。

- 首先，它可以作为单个网络设备进行管理。
- 其次，由于与机箱相同的分段和重组操作，它具有单一控制平面和高度可预测的流量行为。

也就是说，DDC提供了两全其美的方案。DDC具有Clos的灵活性和规模、多供应商以及更好的客户控制。它同时还提供机箱的性能和行为，具有单点管理和单一控制平面，可提高整个互连解决方案的性能。

兼具Clos和机箱优势：Distributed Disaggregated Chassis (DDC)

我们探讨了机箱的性能和管理优势：如果它不是那么僵化，它就具有更好的可扩展性和灵活性来适应不断变化的需求，非常适合人工智能网络。像Clos这样的分布式模型非常适合实现这些属性，并且通过使用标准白盒设计来实施供应商锁定/解锁（或客户控制）。

也就是说，Distributed Disaggregated Chassis (DDC)
是最佳解决方案。我觉得，“开放计算项目”（Open Compute Project）在将其命名DDC时就知道自己的意图了。

English: https://drivenets.com/blog/can-you-really-compare-clos-to-chassis-when-running-ai-applications/