胖树型交换架构中选择柜顶交换（TOR）部署的考虑

作者：Robert Carlson，通信电缆和连接协会（CCCA）数据中心委员会主席

多年来，在数据中心环境使用三层交换机架构是一种常见作法。然而，这种架构无法充分支持大型虚拟化数据中心的低延迟、高带宽需求。由于现在数据中心可随处安置设备，若在三层架构中的两台服务器之间进行数据通信，则需要按照南北通信（即交换机到交换机）模式穿透多次交换层传输，从而增加了传输的延迟时间和网络复杂性。这使许多数据中心转为采用单层或两层的交换矩阵架构。交换机层次越少，经过多层交换机进行通信的需求减少，服务器到服务器的通信就得到改善。

胖树交换矩阵架构也称为枝叶和骨干形式，是目前数据中心最常用的一种交换架构。在胖树交换架构中，数据中心管理人员面对多种配置方案，需要根据应用系统、布线和连接至服务器的接入交换机位置来作出决策。在胖树交换架构中，接入交换机可设置在传统集中式网络配线区域内、列中（MoR）位置或列末（EoR）位置，通过结构化布线连接至服务器。又或者将接入交换机设置在机柜顶（ToR），采用机柜内点对点布线方式连接至服务器。

没有一种理想配置能适合所有的数据中心，新型胖树交换架构的实际运用，要求 CIO、数据中心专业人员和 IT 管理人员能够根据数据中心生态系统的特定需求，进一步审视各种方案的利弊。充分研究、了解各种配置形式、应用系统和布线对可管理性、散热、扩展性和总拥有成本（TCO）的影响，将有利于在从传统三层交换架构转变成新型胖树交换架构时，设施和数据中心管理人员能最终做出最佳的专业决策。

细看各种方案

2013 年 4 月，电信工业协会（TIA）颁布了 ANSI/TIA-942-A-1，这是 ANSI/TIA-942-A 数据中心标准的交换结构布线指南附录。该附录介绍的胖树交换架构由布置在主配线区域（MDA）的互连（主干）交换机和布置在水平配线区域（HDA）和/或设备配线区域（EDA）的接入（分支）交换机构成。各接入交换机通常按照网状拓扑结构，采用光纤连接到各互连交换机（见图 1）。

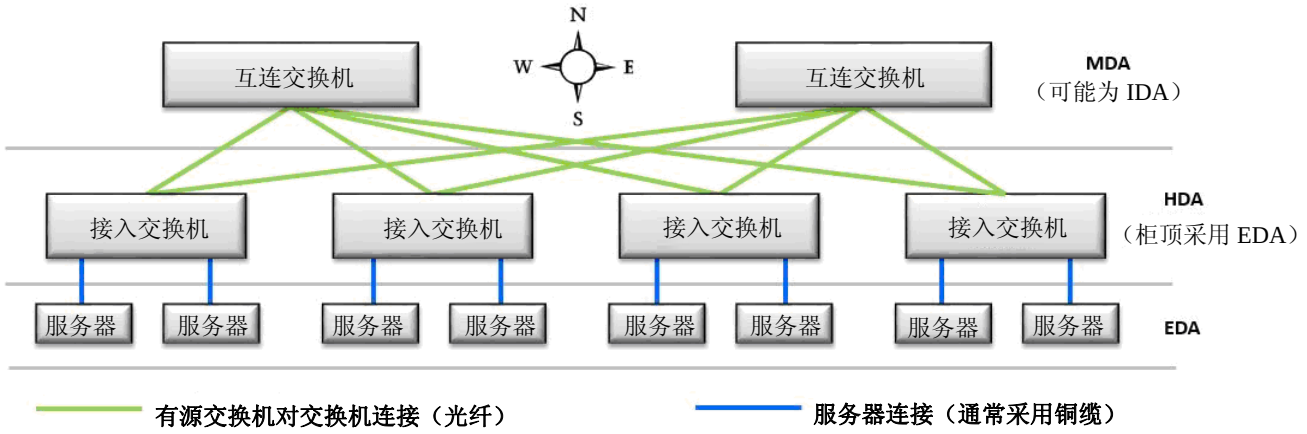


图 1：胖树交换机架构。来源：ANSI/TIA-942-A-1

在胖树交换架构中，可将与服务器和存储设备相连的接入交换机布置在列中（MoR）或列末（EoR）位置，为该列设备提供服务，或者也可布置在独立的专用区域，为多列机柜提供服务（见图 2）。MoR 和 EoR 配置的功能相同，在每一列机柜均专门用于特定目的的数据中心环境中，这两种配置被普遍采用，并可逐列实现增长。本文重点讨论较常用的 EoR 配置。

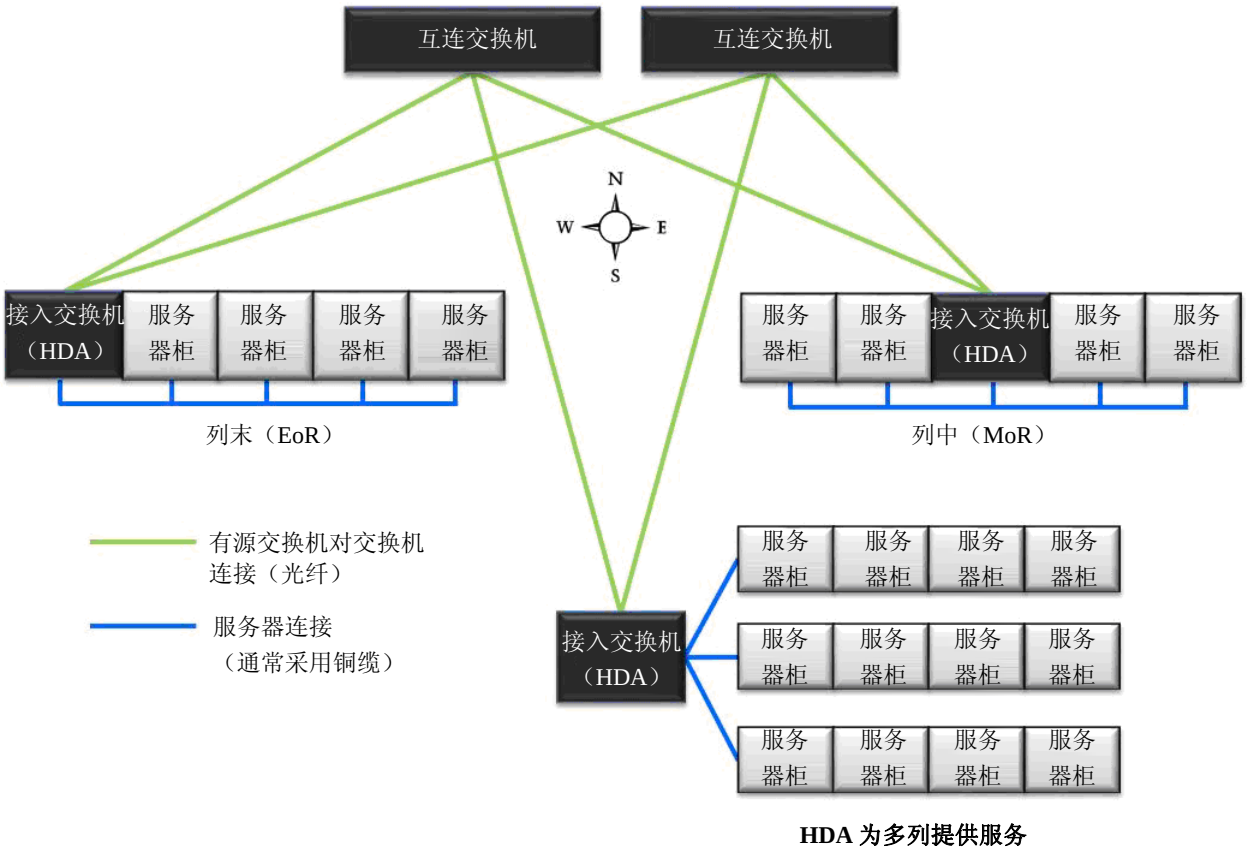


图 2：在胖树架构中，接入交换机（HDA）可设置在 MoR 或 EoR 位置，按列为设备提供服务，或者设置在独立的专用区域，为多列机柜提供服务。

EoR 配置是将接入交换机布置在各列的末端柜，利用结构化布线中无源的配线架作为接入交换机和服务器之间的连接点。位于 EoR 位置的配线架镜像了交换机和服务器的端口（交叉连接），并通过永久链路形式连接至对应的服务器柜内的配线架。通过交叉连接的跳线将交换机和服务器端口相连。

除了将接入交换机布置在 EoR 位置外，也可将其布置在 ToR 位置。在这种方案中，采用光纤布线从 MDA 中的各互连交换机连接到各机柜内的小型（1U 到 2U）接入交换机。各机柜内采用有源的端口扩展器，而不是接入交换机。端口扩展器有时也被称为矩阵扩展器，是对其所属接入交换机的物理扩展。本文以 ToR 交换机来表示设置在 ToR 位置的接入交换机和端口扩展器。

在各机柜内，ToR 交换机以点对点铜缆布线形式直连到该机柜内的服务器，通常会采用预制可插拔小型接口（如 SFP+和 QSFP）的短双轴电缆组件、有源光跳线或 RJ-45 模块化跳线（见图 3）。

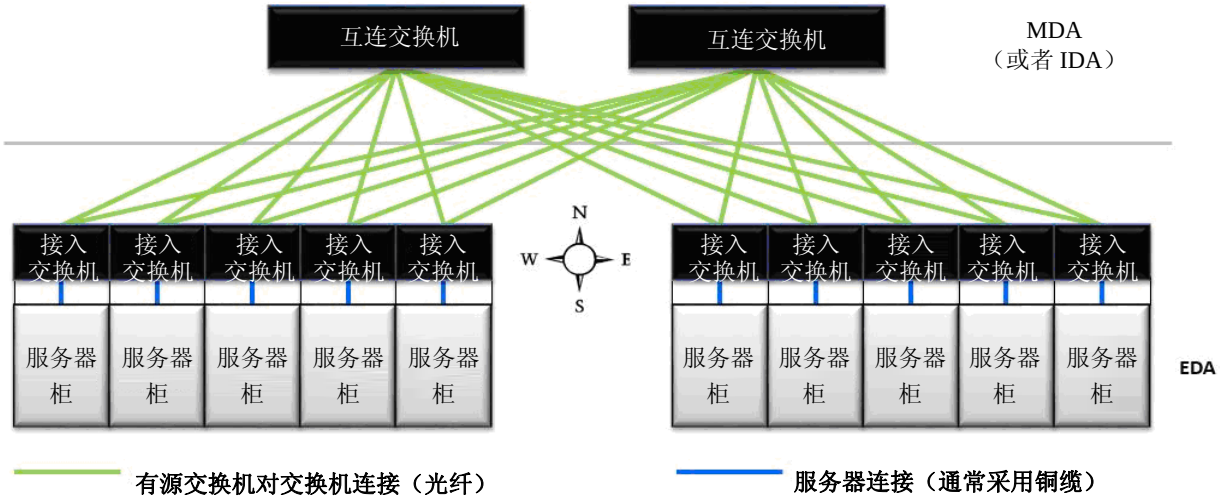


图 3：在 ToR 配置中，各柜顶的小型接入交换机通过点对点连接，直接连接到机柜内设备。来源：TIA-942-A-1

ToR 配置是特别适合密集的机架式（1U）服务器环境，相比于列内通信，机柜内能够快速实现服务器到服务器的连接。对于要求机柜一次性部署和机柜级管理的数据中心，ToR 是理想之选。

采用 ToR 配置时，接入交换机设置在 EDA 内，交换机和服务器之间不需要通过 HDA 和配线区域来连接。实际上，ToR 通常是为了替代和减少结构化布线。然而，结构化布线具有许多优点，包括提高可管理性和扩展性，并且整体上降低了总拥有成本。因此，在胖树交换架构环境中评估 ToR 和结构化布线的配置时，应考虑这些因素。

可管理性

在结构化布线中，有源设备的连接在镜像了设备端口的配线架端口完成，在配线区可完成所有移动、添加和变更的操作（MAC）。通过跳线的简单连接，便可将任一设备端口连接到任何其他设备端口，建立起“any-to-all”配置。

由于 ToR 交换机直接连接同一机柜内的服务器，因此必须在每个机柜内执行所有变更，而不是在便捷的配线区域。由于数据中心规模不同，在各机柜内进行变更会变得较为复杂和费时。想象一下，在成百上千个服务器柜内作变更，而不是在各列 EoR 位置的配线区内完成所有变更。图 4 直观地说明了结构化布线和 ToR 的差异。

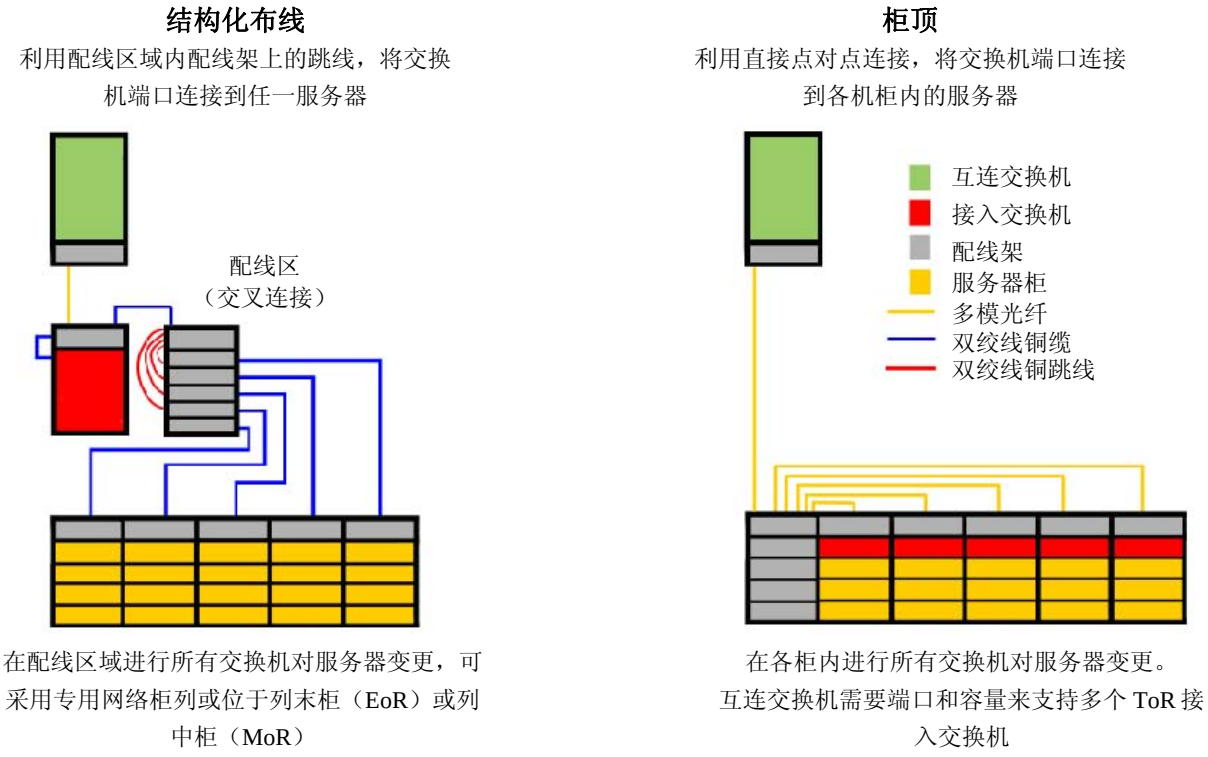


图 4：结构化布线与 ToR 拓扑结构。ToR 省去了便捷的网络变更配线区域。

结构化布线，是用永久链路或固定链路方式，将镜像有源设备端口的配线架连接到配线区域的对应配线架中。由于在配线区域进行所有 MAC，信道的永久部分保持不变，保证有源设备安全不被触及。如图 4 所示，配线区域可设置在完全独立的机柜内，而不需要访问交换机柜。当需要按照独立资源或部门管理交换机和服务器时，结构化布线是理想之选。

ToR 配置不允许将交换机和服务器物理分开到独立机柜内，而且 MAC 时要求访问交换机关键端口。当需要按照应用程序管理服务器组及其对应交换机时，ToR 配置是理想之选。

另一个可管理性考虑因素是跨越多机柜的服务器之间相互通信的能力。虽然 ToR 能够快速实现机架内服务器对服务器的连接，但机柜之间的通信则对交换机与交换机的传输能力有要求。EoR 方式具有这样的优点：同一列的任何两台服务器（非同一柜内）由于连接到同一交换机，因此能够进行低延迟通信。

布线距离的限制也会影响可管理性。对于 ToR 配置，ANSI/TIA-942-A-1 规定点对点布线不宜超过 10 m。而且在无源模式中，ToR 交换机通常使用的 SFP+双轴电缆组件，将交换机和服务器之间的距离长度限制为不能超过 7 米。而结构化布线的信道长度可达到 100 米，因此，在数据中心生命期内，可实现更灵活的设备布局。

散热

由于 SFP+电缆组件的长度较短以及不允许跨机柜配线的数据中心政策，ToR 也限制了设备的布置。用户无法将设备布置在一列或几列机柜内电源和散热最为合理的地方。

例如，如果网络预算不允许再增加一个带 ToR 交换机的机柜来设置新服务器，那么只能将新服务器安装在尚有空闲网络端口的机柜内。而这将可能形成热点，给同一散热区内与其相邻的设备带来不利影响，在某些情况下，还需要附加制冷。而结构化布线配置则避免了上述问题。

在技术上，也可将 ToR 交换机设置在机柜中间或底部，但为了便于使用和管理，通常设置在柜顶。根据 Uptime Institute 的观点，位于机柜上部三分之一范围内的设备的故障率比位于下部三分之二的设备高了三倍。而在结构化布线配置中，无源配线架通常位于上部，因此，能够将设备安装在散热较好的空间。

可扩展性

基于适当的预算和业务模式，ToR 配置可允许以机柜单元扩展，这是某些数据中心偏爱采用 ToR 的原因。与结构化布线相比较，ToR 配置中涉及多个机柜的大范围交换机升级将影响到更多交换机。一台 ToR 交换机的升级虽然会提高连接速度，但仅限于该柜内的服务器。而采用 EoR 结构化布线配置时，一台交换机的升级能够提高同一列上多个机柜内的多台服务器的连接速度。

在可扩展性方面，还应考虑交换机对服务器连接的应用和布线。对于采用结构化布线的 EoR 配置，通常采用标准 6A 类双绞线布线作为电缆介质。在长达 100 m 距离内，6A 类支持 10GBASE-T。10GBASE-T 标准要求在不超过 30 m 的 6A 类或更高性能布线系统上支持短距离（即低功耗）模式。最新的技术发展也使 10GBASE-T 交换机在价格和功耗上迅速下降，从而可与 ToR 交换机媲美。

对于 ToR 配置中的交换机到服务器直接连接，很多数据中心管理人员会选择 SFP+双轴电缆组件，而不是 6A 模块化跳线。虽然这些组件支持低功耗和低延迟，对多端口数的超级计算环境是理想之选，但也有一些缺点必须纳入考虑。

标准 6A 类布线支持自适应机制，但 SFP+电缆组件不支持。自适应是交换机根据连接设备的不同，在各个端口以不同速度自动无缝切换的能力，从而能够按需实现局部交换机或服务器升级。如果无自适应，交换机升级时，需要同时升级与此交换机相接的所有服务器，从而一次性产生全面升级的成本。

几十年来，数据中心管理人员在升级时得益于基于标准的互操作性对现有布线系统的投资影响，而无需考虑采用哪一家供应商的设备。不同于全面支持各种速率和制造商的 BASE-T 交换机产品的 6A 类布线，一些设备供应商可能会要求在其 ToR 交换机上使用昂贵的专有 SFP+电缆组件。虽然这些要求有助于确保在相关电子设备上使用供应商认可的连接线组件，但是专有布线组件无法实现互操作性，也就是在设备升级的同时，可能要进行同步的电缆升级。换句话书，当换用其他牌子交换机时，很可能需要更换所有的 SFP+组件。

有的 ToR 交换机甚至还设计了安全 ID 来校验连至各端口的电缆的供应商，当连接到不支持的供应商 ID 时，即显示故障或禁止端口运行。SFP+电缆组件通常比 6A 类跳线贵，因此将增加升级费用。此外，很多交换机供应商要求的专用连接线平均只有 90 天的质保期。而不同电缆供应商的 6A 类结构化布线则通常能够提供 15 到 25 年的质保期。

设备、维护和能源成本

在 ToR 配置中，每个机柜内至少设置一台交换机，交换机端口总数取决于机柜总数，而不是支持服务器所需的实际交换机端口数。例如，如果有 144 个服务器柜，将需要 144 台 ToR 交换机（若有冗余需求使用主备网络时，需要 288 台）。因此，与使用配线架将接入交换机连至多个机柜内服务器的结构化布线配置相比，ToR 配置所需的交换机数量明显增加。

增加交换机也意味着提高每年的维护费用和能源成本，影响总拥有成本。由于功耗是目前数据中心管理人员最关注的问题之一，这一点尤其值得注意。随着数据中心能耗增加，能源成本不断上升，绿色环保正在成为关注焦点。减少交换机数量有助于降低能源成本，同时促进 LEED、BREEAM 或 STEP 等绿色环保计划。

表 1 以胖树架构的低密度数据中心（144 个机柜）为基准，比较了使用 SFP+电缆组件的 ToR 配置与使用 6A 类 10GBASE-T 结构化布线的 EoR 配置之间的安装、维护和年功耗成本。ToR 配置最终成本比使用 EoR 配置高出 30%。

在本例中，假设每个机柜的平均功率为 5-6kW，每个机柜支持约 14 台服务器。同时基于冗余需求，采用主备双交换机。安装成本包括所有交换机、上行链路、光纤接口卡、光纤主干布线和交换机对服务器铜缆布线。有源设备每年的平均维护成本为设备成本的 15%。年功耗成本以各交换机每天不间断工作的最大额定功率为依据。本例不考虑软件、服务器、机柜和线缆通道的成本。

低密度，144 个服务器柜，每个机柜内 14 台服务器		
物料、功耗和维护	**ToR（SFP+）**	**EoR（10GBASE-T）**
物料成本	$11,786,200	$8,638,300
每年维护成本	$1,655,200	$1,283,100
每年能源成本	$101,400	$44,400
总布线成本（包括在物料成本内）	$1,222,300	$70,300
总拥有成本	**$13,542,800**	**$9,965,800**

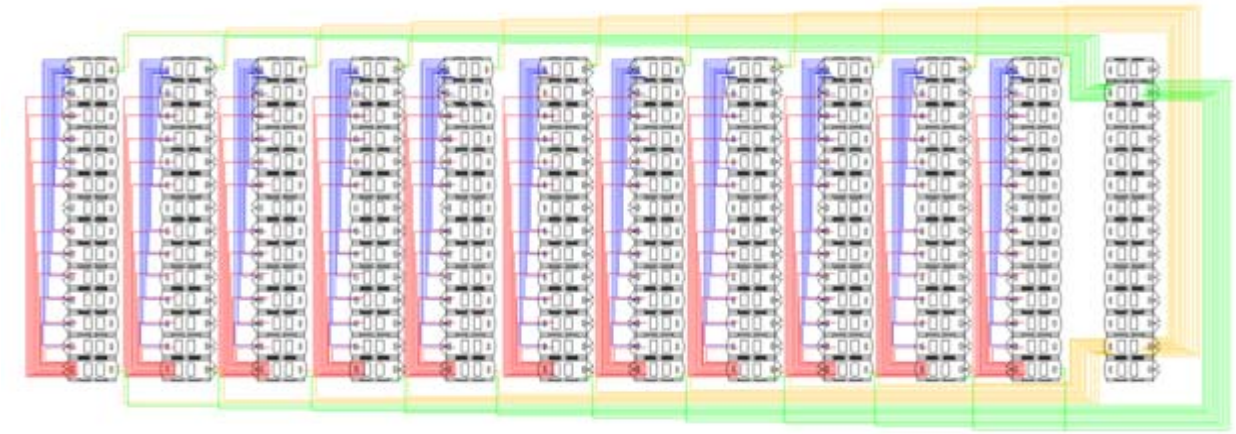
表 1：低密度 SFP+ ToR 与 EoR 结构化布线的成本比较（以印刷时厂商建议零售价为依据，基于 144 个机柜的数据中心案例）

表 2 按照相同的假设，比较了 ToR 配置和 EoR 配置，但基于高密度环境。在该环境下，每个机柜的平均功率为 15-20 kW，以支持 40 台服务器，。在这种情况下，ToR 的总拥有成本仍然比使用 EoR 配置时高出 20%。

高密度，144 个服务器柜，每个机柜内 40 台服务器		
物料、功耗和维护	**ToR（SFP+）**	**EoR（10GBASE-T）**
物料成本	$26,394,000	$21,596,100
每年维护成本	$3,371,900	$2,737,900
每年能源成本	$177,600	$106,700
总布线成本（包括在物料成本内）	$5,123,900	$2,078,200
总拥有成本	**$29,943,500**	**$24,440,700**

表 2：高密度 SFP+ ToR 与 EoR 结构化布线的成本比较（以印刷时厂商建议零售价为依据，基于 144 个机柜的数据中心案例）

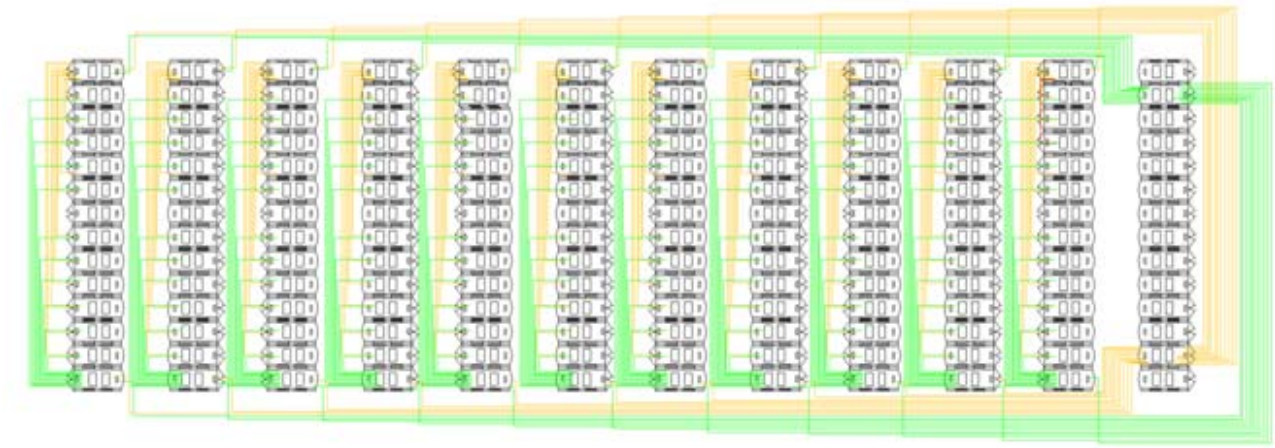
图 5 和图 6 为上述成本案例中使用的 ToR 和 EoR 配置图示。



每列 12 个服务器机柜，核心和 EoR 接入交换机之间为冗余光纤主干（黄色和绿色），EoR 和各服务器柜之间为冗余 6A 类 10GBASE-T 铜缆布线（红色和蓝色）

核心网络交换机（MDA）

图 5：144 个机柜的 EoR 配置



每列 12 个服务器机柜，核心和 EoR 配线区之间以及 EoR 配线区到各服务器柜的 ToR 接入交换机之间，为冗余光纤（黄色和绿色）。点对点铜缆布线被限制在机柜内 ToR 交换机和服务器之间。

核心网络交换机（MDA）

图 6：144 个机柜的 ToR 配置

交换机端口利用率

交换机端口利用率低也意味着总拥有成本较高。在每个机柜仅能容纳 14 台服务器的 5-6 kW 低密度环境下，与 ToR 交换机所提供的 32 个端口相比，服务器对交换机端口的需求较低。如表 3 所示，在表 1 的 144 个机柜案例中，ToR 有 5,184 个未使用的端口，而 EoR 只有 576 个未使用的端口。这相当于多购买了 162 台不必要的交换机，并增加了相应的维护成本和耗电。而使用结构化布线的 EoR 配置时，由于所有有源交换机端口均位于同一机柜内，实际上能够完全利用到这些端口。需要时可通过配线区将交换机端口分配到列内多个机柜里的任何服务器连接。

低密度，144 个服务器柜，每个机柜 14 台服务器	ToR（SFP+）	EoR（10GBASE-T）
未使用的端口总数	5,184	576

表 3：144 个低密度机柜的数据中心案例，采用 ToR 与 EoR 结构化布线的交换机端口利用率（假设每个机柜平均功耗 5-6kW、双网络、14 台服务器/柜）

即使能够向机柜提供足够的供电和散热，以完全满足服务器需求，但采用 ToR 方式时未使用的端口总数仍然明显高于采用 EoR 结构化布线的方案。如图 4 所示，在表 2 使用的高密度 144 个机柜例子中，每个机柜装有 40 台服务器，相当于 ToR 有 6,912 个未使用的端口，而 EoR 只有 224 个未使用的端口。其原因在于为了支持 40 台服务器，各机柜内要求设置两台 32 端口 ToR 交换机，或者在使用主备双网络情况下，设置四台交换机。这相当于每个机柜内有 24 个未使用的端口，或者双网络中有 48 个未使用的端口。在 144 个机柜的数据中心中，未使用的端口数迅速上升。

高密度，144 个服务器柜，每个机柜 40 台服务器	ToR（SFP+）	EoR（10GBASE-T）
未使用的端口总数	6,912	224

表 4：144 个高密度机柜的数据中心案例，采用 ToR 与 EoR 结构化布线的交换机端口利用率（假设每个机柜平均功率 15-20kW、双网络、40 台服务器/柜）

在现实中，真正提高 ToR 交换机端口利用率的唯一办法是将服务器数量限制在不超过各机柜内的交换机端口数量。但是，将服务器数量限制至 ToR 交换机端口数并非总能最有效地利用电源和空间。例如，在支持每柜 40 台服务器的高密度环境下，如每个机柜的服务器数量限制为 32 台（与交换机端口数匹配），则每个机柜将产生 8 个未使用的机架单元，即 144 个机柜的数据中心将产生 1,152 个未使用的机架单元。此外，一旦服务器数量超过可用交换机端口数时，唯一的办法是再增加一台 ToR 交换机（或双网络情况下，需增加两台）。这大大增加了未使用端口数。

无论采用哪种配置，在设计数据中心时，最好要考虑端口利用率，并且能确保有效管理闲置的机架空间和未使用端口。

结论

由于有多种配置方式适用于胖树交换架构，数据中心专业人员和 IT 管理人员需要基于数据中心的特定需求和总拥有成本（TCO）来权衡各种配置的利弊。

没有一种能够适合所有数据中心的布线配置。对于 ToR 配置，在各机架或机柜内布署接入交换机，配备用于交换机到服务器连接的 SFP+电缆组件，是那种对服务器连接的低延迟需求很高，及以机柜级实施部署与维护的数据中心的理想选择。

然而，很多数据中心采用了列末 EoR、列中 MoR 或集中配线等方式来实施 6A 类结构化布线和 10GBASE-T 应用，在可管理性、散热、扩展性、降低成本和提高端口利用率等方面获得了显著的效益。