2026年8月10日

加速器与HBM | AI网络模型 | 核心研究

Rubin Ultra NVL576 架构：快速概览

3分钟

作者 Tony Chien Wega Chu 与 Myron Xie

尽管Rubin Ultra的确切规格和架构设计尚处于变动中，但设计与规格应会很快敲定。在当前阶段，我们已讨论了HBM的规格配置以及在GPU层面提供不同TDP运行的多种SKU。在系统层面，我们还重点介绍了

通过使用NPO技术实现576个Rubin Ultra封装的更大规模扩展。在本说明中，我们将对比Rubin与目前我们所预期的Rubin Ultra规格（后续可能会有变动）。随后，我们将重点讨论旨在实现NVL72跨机架扩展的新型Oberon架构。

Nvidia Roadmap

	2024	2025		2026		2027	
Chip and Package Level							
Generation	Blackwell			Rubin			
Accelerator	B200/ GB200	B300/GB300 (Ultra)	B300 (single die, B300A)	Rubin	LP30/LP35	Rubin Ultra	
GPU TDP (W)	700/1,200	1,100/1,400	600	2,300	~600	1,800/2,600	
Foundry Node	4NP			N3P (3NP)	SF4X	N3P (3NP)	
Logic Die Configuration	2 x Reticle Sized GPU			2 x Reticle Sized GPU, 1 C2C I/O Die, 1 NVLink I/O Die	1 x Reticle Sized LPU	2 x Reticle Sized GPU, 1 NVLink I/O Die	
FP4 PFLOPs - Dense (per Package)	10	15	4.6	35[2]	-	35	
FP8 PFLOPs - Dense (per Package)	5	5	4.6	17.5	1.2	17.5	
FP16 PFLOPs - Dense (per Package)	2.5	2.5	4.6	4	0.6	4	
Memory (per Package)	192GB HBM3E	288GB HBM3E	144GB HBM3E	288GB HBM4	500MB SRAM	192GB HBM4	
HBM Stacks	8		4	8	-	8	
Memory Bandwidth	8TB/s		4TB/s	20TB/s	150TB/s	21TB/s	
Packaging	CoWoS-L			CoWoS-L	FC-BGA	CoWoS-L	
SerDes speed (Gb/s uni-di)	224G			224G Bi-di[3]	112G	224G Bi-di[3]	
System Level and Form Factor							
System	NVL72		NVL16	Rubin NVL8 HGX	Vera Rubin NVL72	Nvidia Groq 3 LPX	Vera Rubin NVL72	Vera Rubin Ultra NVL576
Form Factor	Oberon		HGX	HGX	Oberon	4 racks x 256 Groq 3 LPUs [4]	Oberon	8x Oberon Racks
# of Logical GPUs	72	72	16	8	72	1,024	72	576
# of GPU dies / compute chiplets	144	144	16	16	144	1,024	144	1,152
CPU	Grace		Intel or AMD	N/A	Vera	Intel Granite Rapids	Vera	
# of CPU Sockets	36	36	16	2	36	128 FPGAs 64 CPUs	36	288
Scale up links	Copper Backplane		UBB (PCB)	UBB (PCB)	Copper Backplane	PCB (Intra-node) + Copper Backplane (Inter-node) + AEC/Optics (Inter-rack)	Copper Backplane	Within Rack: Copper Backplane Between Racks: Dragonfly NPO/CPO

来源：SemiAnalysis, Nvidia

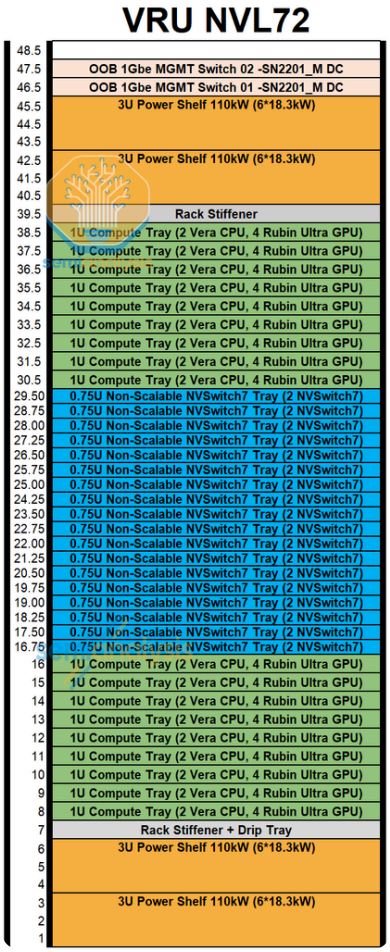
Nvidia Roadmap

	2024	2025		2026				
Scale-Up Networking								
Generation	Blackwell			Rubin				
Accelerator	B200/ GB200	B300/GB300 (Ultra)	B300 (single die, B300A)	Rubin		LP30/LP35	Rubin Ultra	
System	NVL72		NVL16	Rubin NVL8 HGX	Vera Rubin NVL72	1,024 Nvidia Groq 3 LPUs	Vera Rubin Ultra NVL72	Vera Rubin Ultra NVL576
Form Factor	Oberon		HGX	HGX	Oberon	4 racks x 256 Groq 3 LPUs (4)	Oberon	8x Oberon Racks
# of Logical GPUs	72	72	16	8	72	1,024	72	576
# of GPU dies	144	144	16	16	144	1,024	144	1,152
NVLink Generation	NVLink 5			NVLink 6		Groq C2C	NVLink 7	
Scale-Up Topology	One-layer Switched All-to-All			One-Layer Switched All-to-All	One-Layer Switched All-to-All	Dragonfly	One-Layer Switched All-to-All	Direct Connect NPO
Within Rack Scale-Up	Copper Backplane		UBB (PCB)	UBB (PCB)	Copper Backplane	PCB within Tray Copper Backplane between Trays	Copper Backplane	Copper Backplane
Between Rack Scale-Up	N/A			N/A		AEC/Pluggable Optics	N/A	NPO/CPO
Bandwidth per Logical GPU (GB/s uni-di)	900			1,800		1,250	1,800	
Lane Speed (Gb/s uni-di)	200G PAM4			400G (200G PAM4 Simultanous Bidi)		100G	400G (200G PAM4 Simultanous Bidi)	
NVLink Switch Generation	NVLink 5 Switch			NVLink 6 Switch		N/A	NVLink 7 Switch	
Switch Aggregate BW (GB/s uni-di)	3,600			3,600		N/A	3,600	

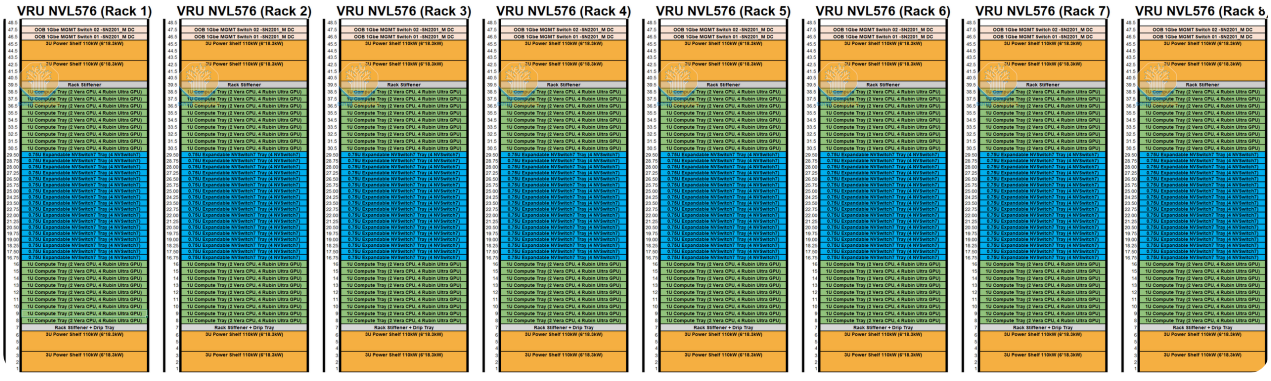
来源：SemiAnalysis, Nvidia

在系统层面，我们可以将该架构分为三个部分：计算托架、NVLink Switch 托架以及机架高度图。Rubin Ultra (Oberon) 的计算托架设计将与 Rubin 大致相同。该

唯一的升级是我们观察到的 Tachyon HPM 板 PCB 层数从 26 层增加到 30 层。此次升级是为了适应从 Rubin 到 Rubin Ultra 更大的 TDP（以支持单颗 Rubin Ultra GPU 最高 2.6KW，尽管我们认为主流将是 1.8KW）。此外，LPDDR5X SOCAMM 将被放置在 HPM 的背面，而不是 HPM 板的正面。



来源：SemiAnalysis, Nvidia



来源：SemiAnalysis, Nvidia

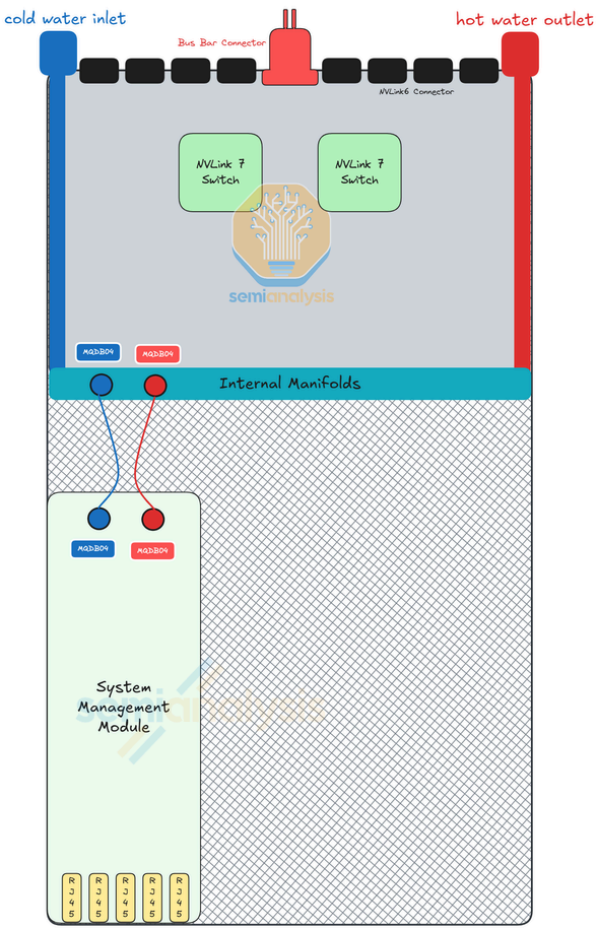
在机架高度方面，变化更为明显，并且
意义重大。目前，Oberon 机架 (10+9+8) 的布局为：顶部布置 10 个计算托架，中部布置 9 个 NVLink Switch 托架，底部布置 8 个计算托架。上述每个托架的高度均为 1U。对于 Rubin Ultra 架构的 Oberon 系统 (9+18+9)，
计算托架的总数维持 18 个不变，但被均匀地分配在机架的顶部和

底部。NVLink Switch 托架的数量增加了一倍，达到每机架 18 个，且单托架高度缩减至 0.75U。此举旨在缩短距离最远的计算托架与最远的 NVLink Switch 托架之间的间距，确保在铜背板上传输的 NVLink 信号能够维持正常的链路驱动能力。此前，计算托架与 NVLink Switch 托架之间的最大距离为 19U；现在，尽管 NVLink Switch 托架数量翻倍，该最大距离仅略微增加至 22.5U。

根据 NVL72 或 NVL576 的纵向扩展规模大小，
NVLink Switch 托盘（代号：Portia）将相应地配置为 NVL72 的不可扩展版本，以及 NVL576 的可扩展版本。在可扩展版本中，目前有 NPO 和 CPO 两个版本正在开发中，

我们认为，考虑到 NPO 相比 CPO 更为成熟的封装形态，NPO 将是率先投入市场的版本。

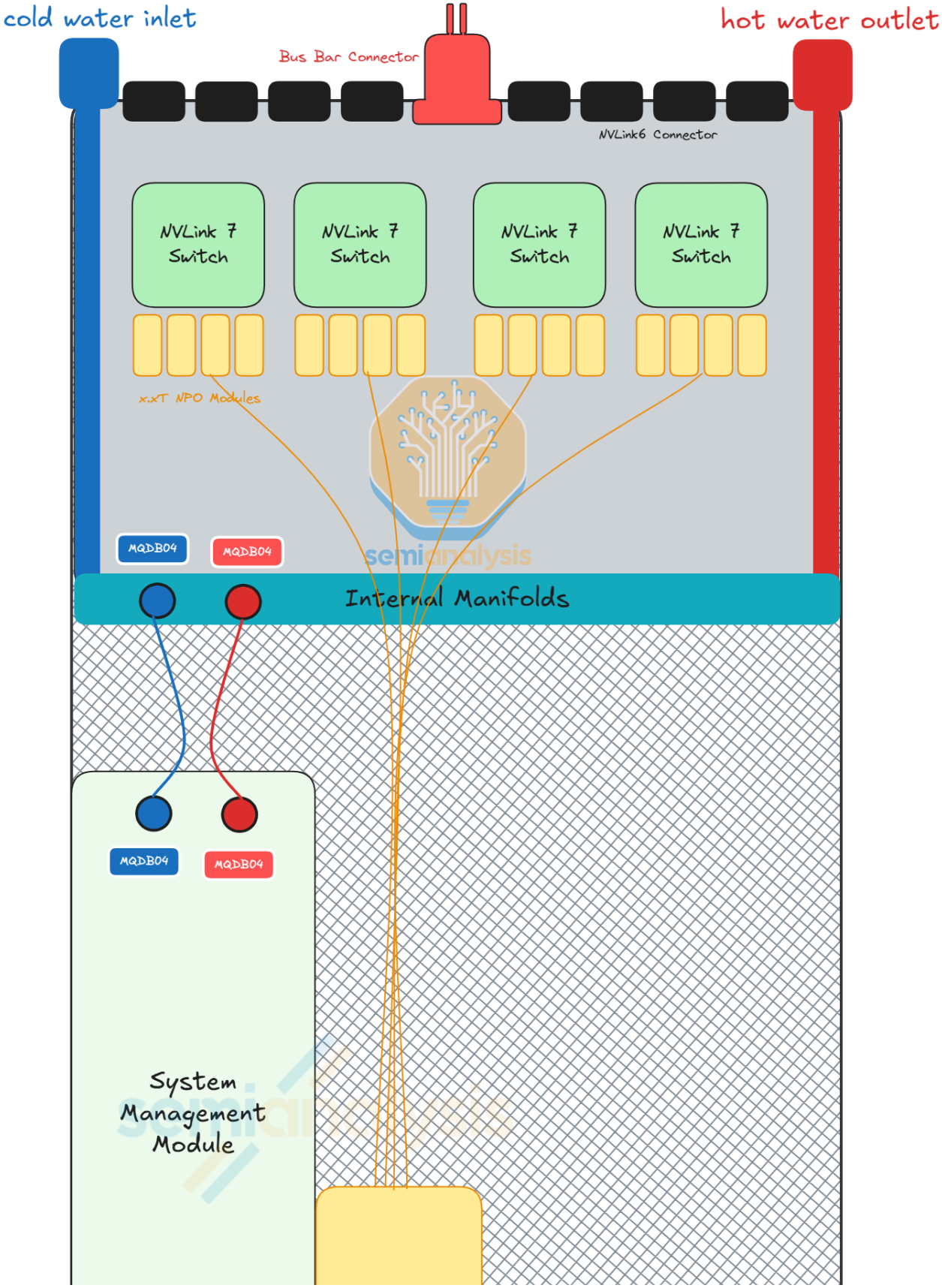
请注意，不可扩展版本和可扩展
版本的机架高度相同，这意味着唯一的区别在于
NVLink Switch 托盘。相同的机架高度基础设施设计使得背板在可扩展和不可扩展版本中保持一致，从而降低了背板端的供应链复杂性。



来源：SemiAnalysis, Nvidia

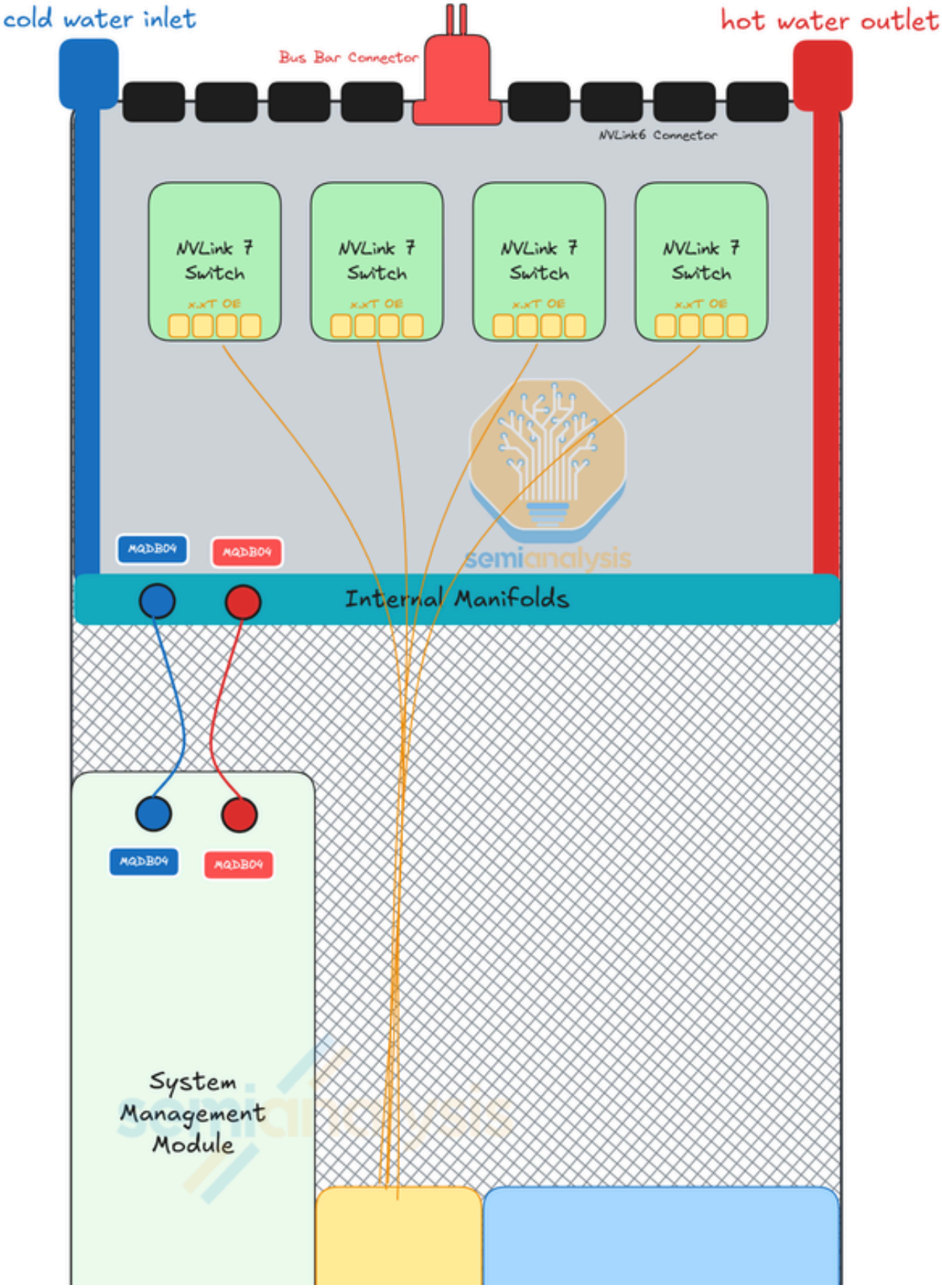
不可扩展的 NVLink Switch 托盘每个托盘将配备 2 个 NVLink Switch ASIC。通过 18 个 NVLink Switch 托盘，每个机架总共配置 36 个 NVLink Switch ASIC，这与 Rubin Oberon 持平。换句话说，不可扩展的 Vera Rubin Ultra Oberon 与 Vera Rubin Oberon 之间的纵向扩展带宽保持一致，且
尽管 NVLink Switch 托盘数量翻倍，但每个机架的纵向扩展 NVLink Switch ASIC 数量依然相同。

VRU NVL576
NVSwitch 7 (NPO) Tray Top View



来源：SemiAnalysis, Nvidia

VRU NVL576
NVSwitch 7 (CPO) Tray Top View



来源：SemiAnalysis, Nvidia

可扩展的 NVLink Switch 托盘分为 NPO 和 CPO 两种变体。对于这两种可扩展版本，NVLink Switch ASIC 的数量比不可扩展版本翻倍至 4 个。这意味着每个机架总共将有 72 个 NVLink Switch ASIC。其目的是提高总带宽，以支持跨机架连接。对于 NPO 版本，NPO 模块将通过插槽安装在 PCB 上，位于 NVLink Switch ASIC 旁边。对于 CPO 版本，每个 NVLink Switch ASIC 将配备 4 个光引擎；然而，这些引擎是不可更换的。

对于新的机架高度，以及不可扩展和可扩展交换机，在硬件和组件方面存在一些细微的影响。新的机架高度对 NVSwitch 的机箱和导轨套件供应商而言是利好消息，因为每个机架中容纳了更多的 NVLink Switch 托盘。背板含量将随 PHD2 连接器升级至 PHD3 而增加，而每个机架的 DP 数量保持不变。Tachyon HPM 的 PCB 层数将增加 4 层，但材料保持不变。NPO 参与者将受益于 NVL576 外形规格，我们已在网络模型说明中对此进行了介绍。最后，NVLink Switch 7 托盘（Portia）的 PCB 将根据交换机版本的不同而有所差异。Portia NPO 版本的 PCB 内容将最为复杂，而非扩展版本的 PCB 虽然也较为复杂，但其面积仅为扩展版本的一半。

免责声明

分析师认证与研究独立性。
本报告中署名的每位分析师在此证明，本报告中所表达的所有观点均准确反映了我们对任何及所有相关证券或发行人的个人观点，且我们的任何报酬过去、现在或未来都不会直接或间接与本报告中的具体建议或观点相关联。SemiAnalysis LLC（以下简称“公司”）是一家独立股票研究机构。本公司不是美国金融业监管局 (FINRA) 或证券投资者保护公司 (SIPC) 的成员，亦非注册经纪交易商或投资顾问。

私人投资披露。
每位参与本报告撰写的分析师和/或作者在此披露，他们可能直接或间接持有本报告所讨论或提及的公司、行业或资产类别的私人投资。此类持仓（如有）均已向第三方供应商进行外部披露。公司要求任何作者或分析师不得让其个人投资头寸影响本报告所表达的观点、意见、分析或结论。读者应意识到，此类持仓的存在（即使已披露）亦可能构成潜在的利益冲突，在评估本报告内容时，建议读者对此进行审慎考量。

研究与信息局限性。
本报告仅供公司合资格的机构或专业客户分发。本报告的内容代表了其作者的观点、意见及分析。本文所含信息不构成财务、法律、税务或其他任何方面的建议。本文呈现的所有第三方数据均获取自被认为可靠的公开渠道；然而，公司对于此类信息的准确性或完整性不作任何明示或暗示的保证。在任何情况下，公司均不对任何此类材料的准确性或更新承担责任，也不对因使用本数据或基于从公司获得的信息或研究而做出的任何投资决策所导致的任何直接、间接、附带或后果性的损失、损害或机会流失（包括利润、收入或商誉损失）承担责任。

本报告或公司发布的任何内容均不应被视为出售或购买任何证券或投资的要约或要约邀请。所收到的任何研究或其他材料均不应被解释为个性化投资建议。投资决策应作为整体投资组合策略的一部分来制定，您在做出任何投资决策前应咨询专业的财务顾问、法律顾问及税务顾问。

严禁复制与分发。
本报告的任何用户不得复制、修改、拷贝、分发、出售、转售、传播、转移、许可、转让或发布本报告本身或其中包含的任何信息。尽管有上述规定，允许接触工作模型的客户可以更改或修改其中包含的信息，前提是仅供该客户个人使用。本报告无意供任何被当地、州、国家或国际法律法规视为非法或以其他方式禁止的用途，亦无意使本公司在相关司法管辖区受到任何形式的注册或监管。

版权、商标、知识产权。
本公司及本报告中包含的任何徽标或标记均为专有材料。未经本公司明确书面同意，严禁使用此类术语、徽标和标记。除非另有说明，本报告页面或屏幕中的版权，以及其中的信息和材料，均为本公司拥有的专有材料。未经授权使用本报告中的任何材料可能违反多项法规、条例和法律，包括但不限于版权、商标、商业秘密或专利法。

ETF 及监管状态声明。
本公司可能不时与资产管理公司、发行人、发起人、指数提供商或其他金融机构合作，共同开发、推出或支持交易所交易基金（“ETF”）及其他投资产品，包括提供研究、市场分析、指数相关研究或其他与研究相关的服务。本公司并不

发起、认可、销售或推广任何 ETF，且对投资此类 ETF 的建议性不作任何明示或暗示的陈述或保证。就本文所述服务而言，本公司不担任任何投资者、基金、发起人、发行人或其他方的受托人。本公司并未在美国证券交易委员会或任何州证券监管机构注册为投资顾问或豁免申报顾问。本公司的角色仅限于提供研究及相关服务，不提供个性化投资建议、投资组合管理、经纪服务或其他需要注册为投资顾问的服务。

// 相关研究

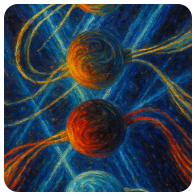
继续阅读。

2026年8月10日 核心研究

TPU 繁荣昌盛：别因为 Gemini 的落幕而哭泣，要为 TPU 的诞生而微笑

核心研究独家报道

作者：Konrad Wang 和 Tony Chien



2026年8月7日 AI网络模型

互连史诗对决：VCSELs 与 MicroLEDs

作者：Julien Martin-Prin 和 Daniel Nishball

2026年8月7日 加速器与HBM 核心研究 内存模型

HBM 定价上更倾向于三星而非 SK 海力士

多垂直领域简报

作者：Timothy Chang、Nick Doyle 和 Dylan Patel