
# InfiniBand 常见问题解答

**1.3** 版

www.mellanox.com

注释：
此硬件、软件或测试套件产品（“产品”）及相关说明文件都由 MELLANOX TECHNOLOGIES 按原样提供，可能出现各种错误，仅用于协助客户在指定解决方案中测试使用产品的应用。客户的生产测试环境尚未达到 MELLANOX TECHNOLOGIES 设定 DESIGNATED SOLUTIONS，不完全符合产品和/或系统的使用要求。因此，MELLANOX TECHNOLOGIES 不能也不会保证或担保产品能够以其最高质量来运行。本公司不提供任何明示或暗示的保证，包括但不限于对于适销性、是否适合特别目的与非侵权的默认保证。在任何情况下，不管是否告知过出现此类损伤的可能性，无论是否在合同范围内，MELLANOX 不会向客户或任何第三方承担在使用产品和相关说明文件的过程中产生的任何形式的直接、间接、特别、惩戒性或后续损伤责任（包括但不限于，支付替代产品或服务的采购费用；使用、数据或盈利的损失；或业务中断）、理论责任、限制责任或侵权责任（包括过失或其他原因）。

# 目录

## 目录

# 1 什么是 InfiniBand (IB)?

InfiniBand 是一种可在处理器节点之间，以及处理器节点与 I/O 节点（例如磁盘或存储器）之间提供基于交换的点对点双向串行链路网络的
网络通信协议。每一条链路只有一个连接至链路两端的设备，如此才能很好的定义和控制每个终端的传输控制（发送和接收）特点。

InfiniBand 通过交换机在节点之间直接创建一个受保护的、私有的通道，借助远程直接内存访问 (RDMA) 与由 InfiniBand 适配器管理和执行的发送/接收卸载，无需占用 CPU 即可为数据和消息移动提供便利。适配器一端通过 PCI Express 接口连接至 CPU，另一端通过 InfiniBand 网络端口连接至 InfiniBand 子网。和其他网络通讯协议相比，这种方式优势明显，包括带宽较高、延迟较低与扩展性增强。

1999 年成立的 InfiniBand 贸易协会 (IBTA) 制定、维护和优化了 InfiniBand 规范，并负责商用 InfiniBand 产品的合规性与互操作性测试。通过其发展路线图，IBTA 更为积极地推动了与其他互联解决方案相比性能更高的产品开发，以确保能够设计一个面向 21 世纪的架构。Mellanox 很荣幸成为 IBTA 指导委员会活跃成员。

有关 InfiniBand 的更多信息请访问 https://cw.infinibandta.org/document/dl/7268

# 2 InfiniBand 与传统网络协议的不同之处在哪儿?

InfiniBand 设计用于实现最高效率的数据中心实施。它与生俱来就支持服务器虚拟化、叠加网络和软件定义的网络(SDN)。

InfiniBand 采用一种以应用程序为中心的方式来传送信息，从而寻找最小阻力的路径在点之间传递数据。与更侧重于以网络为中心的通信方式的传统网络协议（例如 TCP/IP 与光纤通道）相比，这种模式有很大不同。

直接访问意味着应用程序无需依赖操作系统来传递消息。在传统的互连模式下，操作系统是共享网络资源的唯一所有者，这就意味着应用程序无法直接访问网络，而是必须依赖操作系统将数据从应用程序的虚拟缓冲器传送到网络堆栈与线路，而接收端的操作系统也必须有类似的活动，只不过方向刚好相反。

相比之下，InfiniBand 绕过网络堆栈在两端的应用程序之间建立起直接的通信通道，从而避免了操作系统的介入。InfiniBand 的明确目标就是为应用程序直接与另一个应用程序或存储器通信提供消息传递服务。一旦建立起这种联系，InfiniBand 架构的其他部分就是为了确保这些通道能够携带各种规格的消息至横跨巨大物理距离的虚拟地址空间，并且保证隔离与安全性。

# 3 InfiniBand 和 TCP 的区别在于哪些方面？

InfiniBand 是为硬件实施而架构的，而 TCP 架构主要是考虑软件实施。因此，和 TCP 相比，InfiniBand 是一种更轻量型的传输服务，无需对数据包重新排序，因为较低层级的链路层可提供有序的数据包传递。传输层仅需要检查数据包顺序并按照顺序传递数据包。此外，由于 InfiniBand 可提供基于信用的流控制(发送方节点不会发送超出由链路的相反方接收缓冲器公布的“信用”额度的数据），传输层不会要求 TCP 加窗算法之类的丢包机制来确定未送达数据包的最佳数量。这就可以实现能够以 56Gb/s（很快就能达到 100Gb/s）的速率将数据传输至应用程序的高效产品，且延时极低，CPU 使用率可以忽略。

# 4 InfiniBand 是一种严格意义上的 I/O 网络吗？

不，InfiniBand 可以提供更多功能。在最低层，InfiniBand 可作为可伸缩的 I/O 互连来提供高性能低延时的可靠交换机网络。当然，InfiniBand 还可为较高层提供能够实现应用程序集群化、虚拟化和存储区域网络（SAN）的功能。

# 5 InfiniBand 是一种分层协议吗？

是的。InfiniBand 规格按模块化层来定义协议，大致基于 OSI 7 层模型并包含 1 到 4 层。此规格定义了给定层与上层及下层之间的接口。因此，最低物理层仅能连接到上层链路。InfiniBand 链路层定义了连接下级物理层的一个接口，以及连接上级网络层的另一个接口。

# 6 InfiniBand 有哪些优势？

InfiniBand 相对于其他互连技术的主要优势包括：

- 吞吐量更高 – 每服务器和存储器连接 56Gb/s（很快就将达到 100Gb/s），而以太网和光纤通道最大只能达到 40Gb/s
- 延时更低 – RDMA 零拷贝网络可降低操作系统开销，使数据能够通过网络快速移动
- 伸缩性更强 – 通过在相同交换机组件的基础上简单添加额外交换机，InfiniBand 理论上可以适应无限规模的平面网络
- CPU 效率更高 – 借助数据移动卸载，CPU 可以将更多的计算周期用于应用程序上，从而减少运行时间，增加每天的任务数量
- 投资回报率（ROI）更高– 以有竞争力的定价获得更高的吞吐量和 CPU 效率等于是以更低的每端点成本获得更高的生产率

# 7 InfiniBand 的可用数据速率为多少？

InfiniBand 串行链路可以在不同的信号传输速率下运行，然后可以聚合在一起以获得更高的吞吐量。原始信号传输速率都搭配一个编码方案，从而实现有效的传输速率。编码将铜线或光纤之间发送数据的错误率降至最低，并增加了某些开销（例如，每 8 比特数据需要传输 10 比特）。

典型的实施做法为聚合 4 链路单元（4X）。当前 InfiniBand 系统可提供如下吞吐量速率（每链路每个方向）：

***表 1:InfiniBand 数据速率***

| 名称 | 缩写 | 原始信号传输速率 | 应用的编码 | 有效数据速率 | 聚合 (4x) 吞吐量 |
|---|---|---|---|---|---|
| 单倍数据速率 | SDR | 2.5 Gb/s | 8b/10b | 2 Gb/s | 8 Gb/s |
| 双倍数据速率 | DDR | 5 Gb/s | 8b/10b | 4 Gb/s | 16 Gb/s |
| 四倍数据速率 | QDR | 10 Gb/s | 8b/10b | 8 Gb/s | 32 Gb/s |
| 十四倍数据速率 | FDR | 14.1 Gb/s | 64b/66b | 13.64 Gb/s | 54.5 Gb/s |
| 增强型数据速率 | EDR | 25.8 Gb/s | 64b/66b | 25 Gb/s | 100 Gb/s |
| 高数据速率 | HDR | 51.6 Gb/s | 64b/66b | 50 Gb/s | 200 Gb/s |
| 下一代数据速率 | NDR | TBD | TBD | TBD | TBD |

有关 InfiniBand 数据速率的更多信息，请访问：
http://www.infinibandta.org/content/pages.php?pg=technology_overview

# 8 什么是 RDMA?它有哪些好处?

InfiniBand 使用远程直接内存访问 (RDMA) 作为从通道的一端向另一端传送数据的方式。RDMA 是通过网络直接在应用程序之间传送数据的能力，无需操作系统参与，与此同时收发两端都只消耗可忽略的 CPU 资源（零拷贝传送）。另一侧应用程序只需要简单地从内存直接读取早已成功传送的消息即可。

这就降低了 CPU 开销，提高了网络快速移动数据的能力，并允许应用程序更快的接收数据。从源系统到目的地系统传输给定数量数据的时间间隔就是所谓的延时，延时越低，应用程序完成任务就越快。

***图 1:传统互连***

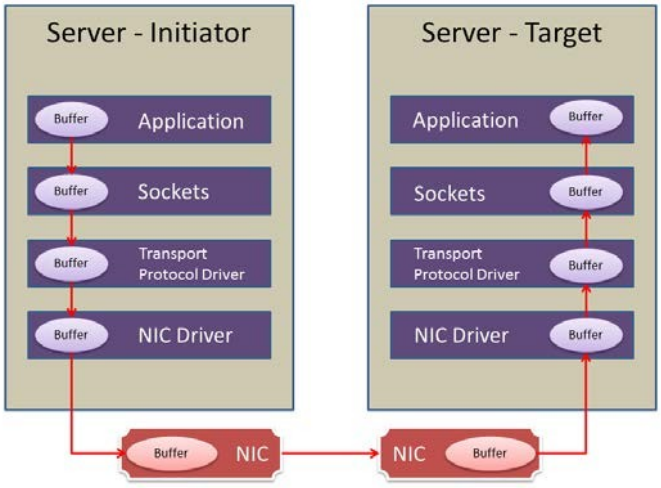
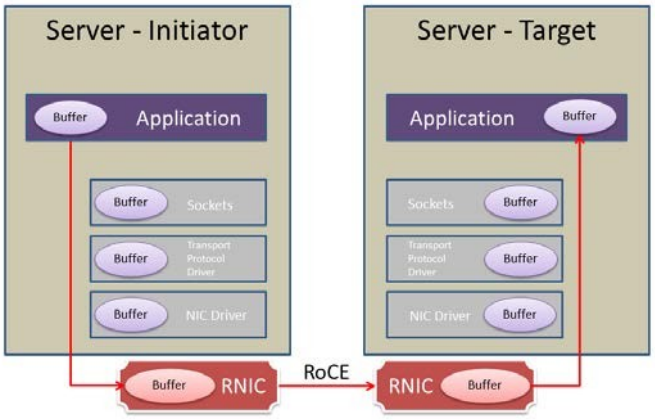



***图 2:RDMA 零拷贝互连***



Mellanox FDR InfiniBand 已实现 0.7 微秒的极低延时，这无疑是数据传输可用的最低延时。

有关 RDMA 的更多信息请访问：http://www.mellanox.com/blog/tag/rdma/

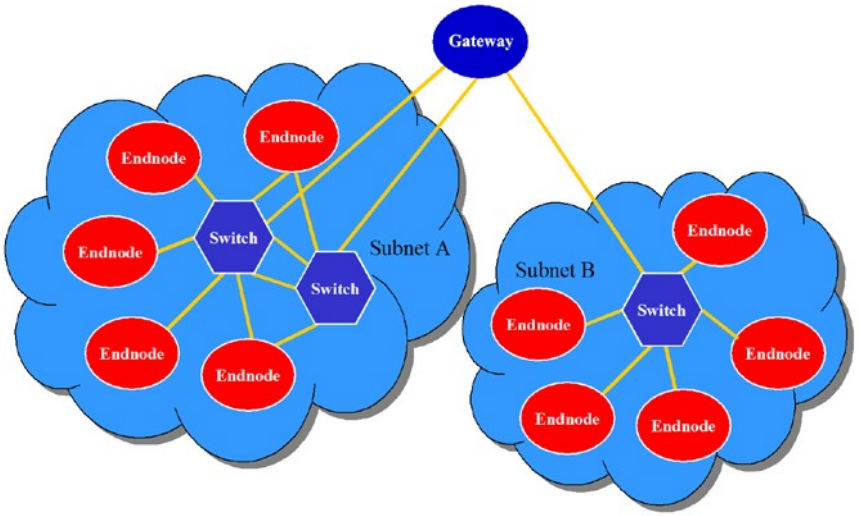
# 9 InfiniBand 架构的主要元素是什么？

InfiniBand 网络的基本构建模块有：

- 主机通道适配器 (HCA)
- 交换机
- 子网管理器 (SM)
- 网关

每个终端节点必须有一个主机通道适配器来设置和维护与主机设备的连接。交换机包含多个 InfiniBand 端口，将数据包从一个端口转发到另一个端口，以便在一个子网内继续传输数据包。路由器用来将数据包从一个子网转发到另一个子网（如有必要）。子网管理功能是通过一个软件定义的网络来处理的,此类网络一般通过业界标准的开放接口控制网络的物理元素及引入的流量控制功能。

***图 3:基本 InfiniBand 架构***




有关 InfiniBand 架构的更多详细信息,请访问：http://www.mellanox.com/related-docs/solutions/deploying-hpc-cluster-with-mellanox-infini band-interconnect-solutions.pdf

## 9.1 什么是主机通道适配器（HCA）？

主机通道适配器（HCA）是桥接 InfiniBand 线路和主机系统总线的一个接口卡或控制器。每个终端节点必须有一个 HCA，用于设置和维护主机设备和网络中其他实体间的链路。

HCA 提供与其他设备的端口连接。 HCA 可连接至另一个 HCA、一个目标设备或一个交换机。

## 9.2 什么是交换机？它在 InfiniBand 网络中如何工作？

交换机用于以物理方式连接网络内设备，并将传入数据向其目的地转发。交换机有多个端口，用于跨线缆处理数据并将数据转发至预期的特定设备，同时在网络内调节流量。

在 InfiniBand 网络中，交换机是架构的一个关键部分。事实上，InfiniBand 之所以称为“交换式网络”或“基于交换的互连”，是因为在转发流量时，从一个端口到另一个端口有一个逻辑连接，类似于老式的电话总机。随着更多的交换机添加至系统中，设备间甚至还会有更多可能的路径。

## 9.3 什么是子网管理器 (SM)？

子网管理器 (SM) 是一种软件实体，可配置本地子网并确保其持续运行。子网管理器在每个端点之间设置主要和次要路径来预先决定流量转发的路径，使数据在最短的时间内到达。子网中必须要有至少一个子网管理器，以便管理所有交换机和路由器设置，并在链路停机或新链路上线时重新配置子网。子网管理器可以驻留在子网内任意设备上。

子网内所有设备必须包含一个专用的子网管理代理程序（SMA），子网管理器使用该代理程序与不同的 InfiniBand 组件通信。

子网内可以有多个子网管理器，但是任何时间都只有一个为活跃状态。非活跃状态的子网管理器被称之为备用子网管理器，负责保存活跃子网管理器的转发信息，并核实活跃子网管理器是否在运行中。如果活跃子网管理器停机，备用子网管理器会接手，确保整个网络继续运行。

软件定义的网络 (SDN) 已成为子网管理的主要方式，SDN 不仅可以控制网络设备及设备之间的数据流量，还可以改善网络灵活性和伸缩性。

## 9.4 路由器是不是 InfiniBand 网络中的必要组成部分？

路由器用于分离的计算机网络之间设备的物理连接。数据抵达路由器时，路由器会读取数据包中的地址信息，并确定其最终目的地。

在以太网网络中，由于洪泛机制和生成树，大型子网的效率不高。因此需要路由器来桥接不同小型子网。

这样的问题在 InfiniBand 中不存在，多达约 40000 个节点的大型子网也能够高效率运行。无需分成小型子网，所以路由器也不是必需的。

但是，随着对数据中心需求的不断增长，如果需要超过 40000 个节点的子网，则可以提供诸如 Mellanox 开发的 Switch-IB™ 等 InfiniBand 路由器技术。

## 9.5 什么是网关？它在 InfiniBand 网络中如何工作？

网关是子网中桥接两个协议的设备。例如，它允许 InfiniBand 群集访问以太网网络管理或存储器接口。这就使得公司可以为运行其他协议的当前旧有装置实施一个配置专用界面的 InfiniBand 网络。连接 InfiniBand 群集与以太网网络不需要网关。部分或全部计算节点可以通过利用以太网适配器加上 InfiniBand HCA 或借助虚拟协议互连®使用 HCA 来访问以太网。

有关 InfiniBand 网关的更多信息，请访问：
http://www.mellanox.com/page/gateway_overview

# 10 什么是 Virtual Protocol Interconnect®？它与 InfiniBand 有什么关系？

Virtual Protocol Interconnect® (VPI) 是 Mellanox 的一种技术，可为客户提供最佳可用互连，与此同时确保可承受的伸缩性。VPI 允许 HCA 内的任意端口或交换机运行 InfiniBand 或以太网，并且如有需要还可在二种协议之间切换。对于服务器和存储系统来说，这就可以实现最高的连接灵活性。而对于交换机网络来说，这相当于创建了一个全新的完美网关，从而可以实现 InfiniBand 和以太网网络与集群之间的集成。

Mellanox VPI 可同时提供端口灵活性与财务灵活性。您可以考虑增长情况按需增减端口，而不是按增长付费。此外，Mellanox VPI 可提供 InfiniBand 的所有好处，与此同时还可维持与当前以太网网络的轻松连接。最重要的是，无需付出性能代价即可实现，因为 Mellanox 在其 InfiniBand 和以太网服务中都可以确保最高的带宽和最低的延时。

有关 InfiniBand VPI 的更多信息，请访问：http://www.mellanox.com/related-docs/case_studies/CS_VPI_GW.pdf

# 11 什么是 LID、GID 和 GUID？

子网中所有的设备都有一个本地标识符（LID），一个由子网管理器分配的 16 位地址。子网内所有发送的数据包都使用 LID 作为链路层级转发和交换数据包的目的地地址。LID 允许单个子网内存在最多 48,000 个终端节点。在重新配置子网时，子网内不同端点会分配到新的 LID。

不同子网之间的路由建立在全局标识符（GID）的基础上，GID 是在 IPv6 地址基础上改编的一个 128 位地址， 它确保了 InfiniBand 最基本的无限伸缩性。GID 可确定终端节点、端口、交换机或多播组。

全局唯一标识符 (GUID) 是子网内所有元素的 64 位定义，包括机架、HCA、交换机、路由器和端口。GUID 永远不会发送变化，并用作创建 GID 地址的一部分。

GID、GUID 独立于 LID，不受子网重新配置的影响。

# 12 InfiniBand 支持 IP 流量吗？什么是 IPoIB？

因特网协议 (IP) 数据包可以通过借助一个网络接口将 IP 数据包封装在一个 InfiniBand 数据包内，然后通过 InfiniBand 界面发送。这就是所谓的 IPoIB。在 InfiniBand 网络安装了必要的驱动程序以后，会使用分区键值 (PKEY)为每个端口创建一个接口，然后跨 InfiniBand 网络无缝传输 IP 数据包。

有关 IPoIB 的更多信息，请参阅 Mellanox OFED 用户手册 或 Mellanox WinOF VPI 用户手册。

# 13 什么是可靠与不可靠传输模式？

数据包从一个节点向另一个节点传输时，可以通过可靠或不可靠方式来传输。正如其名，可靠传输不会有数据包的丢失或重复，数据包必须按照顺序传送，接收节点向发送方提供确认（肯定或否定），以确认数据包是否正确接收。通过确保借助硬件协议向远程应用程序传送数据，应用程序本身就无需承担此种责任。

而不可靠传输则不会提供此种确认，也不会尽力去按照顺序来传送数据包（使用不可靠传输时也被称之为数据报文）。

可靠传输要求保留额外资源来确保数据包能够正确收发，以及处理确认。因此，有些时候，数据中心为了将这些资源用于他处，可能会放弃可靠连接，而使用不可靠的数据报。

IPoIB 既可以在适用于不可靠传输的不可靠数据报（UD）模式上运行，也可以在能够保证可靠传输的连接模式（CM）上运行。借助单个命令即可在运行期间在这个两种模式之间切换。默认的 InfiniBand 传输模式为 UD，但是也有可能添加一个脚本来将新接口自动配置为 CM 模式，以实现可靠传输。

# 14 IPoIB 支持绑定吗？

绑定指的是合并两个端口供单个应用程序所用的程序，可为发送数据提供更大的灵活性。IPoIB 接口的绑定仅适用于 Linux，与以太网接口的绑定流程相同，也就是说可以通过 Linux 绑定驱动程序进行绑定。Windows 当前并不支持绑定。

# 15 InfiniBand 支持多播吗？

交换机可被配置用于转发单播数据包（至单个位置）或多播数据包（至多个设备）。子网管理器负责管理这些可完全热转换的连接，并且还可包含发送至子网上所有系统或这些系统的子网。InfiniBand 的高带宽可为此类多播功能提供骨干网，无需第二个互连链路。

# 16 InfiniBand 支持服务质量吗？

服务质量控制 (QoS) 是一种网络功能，可为应用程序提供不同优先度，并保证数据流向这些端点过程中保证一定等级的性能。

InfiniBand 通过创建虚拟通道（VL）的方式来支持 QoS。这些 VL 都是共享一个单独物理链路的逻辑通信链路。每个链路可支持最多 15 个标准 VL（指定为 VL0 到 VL14）与一个管理通道（指定为 VL15）。
VL15 优先度最高，而 VL0 优先度最低。

服务级别（SL）可用于定义数据包，以确保其 QoS 级别。SL 为每个链路提供要求的通讯优先度。子网管理器会为每个交换机或路由器与通信路径之间设置一个映射表，以将 SL 转换为 VL，确保支持该链路的 VL 内适当的优先度。

在主机端，会为每个流（队列对）分配 QoS 参数，为每个端点提供 1600 个万 QoS 级别。

当应用程序的带宽需求达到定义的 QoS 上限时，主机应用程序使用的资源将无法再增加。降低已达到定义上限的应用程序的 QoS 限制可为主机释放资源，而主机的其他工作负载则将获益。这对于流量共享来说有很高的价值。

有关服务质量的更多信息，请参阅 Mellanox OFED 用户指南或 Mellanox WinOF VPI 用户指南。

# 17 InfiniBand 是不是一个无损网络？

无损网络指的是一般情况下数据包不会丢失的网络。以太网被认为是一个有损网络，因为经常丢包。TCP 传输层可以检测到丢失的数据包并会做出调整。

相反，InfiniBand 采用链路层面的流控制以确保网络中不会丢失数据包。这种无损流控制可带来极为高效的数据中心内带宽使用，令 InfiniBand 非常适合数据中心间远程通信。

# 18 InfiniBand 如何处理安全问题？

InfiniBand 的适配器为高级虚拟化功能—例如网络功能虚拟化（NFV）—提供支持。与 SDN 优化的 InfiniBand 交换机搭配使用后，就有可能构建最高效的可伸缩高性能网络。在数据移动和分析的速度方面，InfiniBand 解决方案比任何其他解决方案都要快，这就为数据中心经理提供了收集网络流量信息来分析流量行为（例如异常现象的检测）并确保其安全的能力。此解决方案为构建防火墙和入侵检测解决方案的基础架构提供了可能。此外，通过 InfiniBand 解决方案实现的网络流量控制非常灵活，可实时再编程。

如需采用 SR-IOV 适配器的基于 SDN 的安全解决方案的示例，请参阅以下博文。

# 19 基于信用的流控制的工作原理是什么？

流控制用于管理链路间的数据流，以保证无损网络。链路（虚拟通道）的每一个接收端都为发送设备提供信用，以指定在数据无损的情况下能够接收的数据量。有一个专用的链路数据包管理设备间的信用传递，以更新接收设备可以接收的数据包数量。除非接收端发布信用，说明有足够的缓冲空间来接收整个消息，否则不会传送数据。

# 20 InfiniBand 有生成树吗？

生成树被视为以太网优势劣势最为极端的一个元素。在节点之间存在环路时，它允许以太网网络诊断并禁用冗余或并行链路，以阻止数据包重复发送。

在 10 到 15 年前，这一功能非常有用，而现在生成树被视为一种“限制因素”，它限制网络工程师构建高性能网络的能力，因为它对节点之间的并行路径的数量设置了限制。

而另一方面，Infiniband 却通过一个可以看见网络内所有路径的集中代理程序来管理流量，以此来实现所有端点间全带宽的大型链路网络。今天，软件定义的网络（SDN）承担此项集中子网管理任务，配送和分配流量以充分利用端点之间的所有并行链路。显然这种做法更好，因为 SDN 可在连接设置时确定配置内的可能路径，所以这种管理对性能没有任何影响。

# 21 什么是 Verbs？

Verbs 是应用程序从 InfiniBand 的消息传输服务请求一个操作的方法。Verbs 集是指应用程序与 InfiniBand 网络交互所使用的各种操作的语义描述。Verbs 在 InfiniBand 软件传输接口规范中有完整的定义。

这些 Verbs 是指定应用程序所使用 API 的基础，但是 InfiniBand 架构并不会定义 API。其他组织，例如 OpenFabrics Alliance，会提供实施 Verbs 以便和 InfiniBand 硬件无缝搭配使用的完整 API 和软件集。

有关 InfiniBand Verbs 使用的更多信息，请访问：

https://cw.infinibandta.org/document/dl/7268

http://www.mellanox.com/related-docs/prod_software/RDMA_Aware_Programming_user_manual.pdf

# 22 我如何监控 InfiniBand 网络的带宽、拥塞和健康状况？

OpenFabrics Enterprise Distribution (OFED) 是一个支持 InfiniBand 网络的软件驱动程序、核心代码、中间件和用户层面接口的开源堆栈。适用于 Linux 的 Mellanox OFED 与适用于 Windows 的 Mellanox OFED (WinOF) 包括可监控 InfiniBand 网络健康状态的各种诊断和性能工具，包括监控传输带宽与网络内拥塞监控。

Mellanox 还提供 Unified Fabric Manager® (UFM®) 软件，一种用于管理 InfiniBand 计算环境的强大平台。UFM 可让数据中心操作员有效配置、监控和操作网络，与此同时提升应用程序性能，并确保网络在任何时间都能运行。