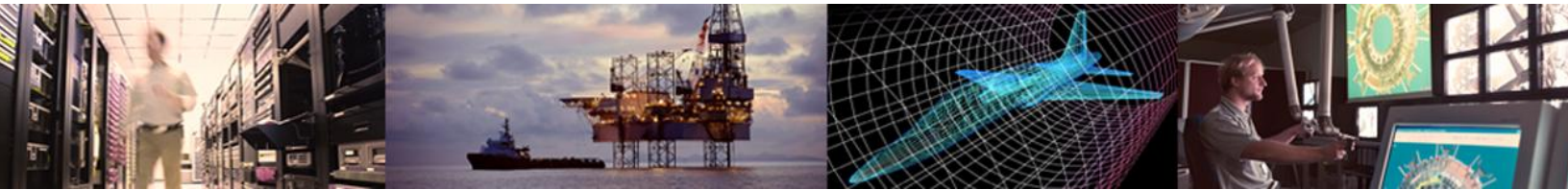


Technology Brief

Solarflare OpenOnload vs. RDMA over Converged Ethernet (RoCE)

Solarflare OpenOnload provides superior ultra-low latency for high-performance networking



SOLARFLARE® OPENONLOAD® PROVIDES PERFORMANCE WITHOUT PAIN

OpenOnload is a dynamically linked library which transparently accelerates all POSIX socket API calls, while ensuring application compatibility, easy deployment, and simplified management. OpenOnload achieves ultra-low latency while maintaining full compatibility with standard 10G Ethernet (10 GbE) networks, protocols, and applications, without requiring the added costs and complexities of special networking equipment, software modifications, and special management tools. As a result, the deployment of OpenOnload adds no incremental CapEx and OpEx to the IT budget.

ULTRA-LOW LATENCY

OpenOnload's kernel (OS) bypass technology cuts latency by 50%, or more, and by moving packets directly between the network wire and application buffer, can eliminate kernel overheads, including memory copies, context switching, and interrupts. OpenOnload also dramatically reduces streaming latency and jitter, enabling far greater message rates and superior application performance at very low CPU utilization. OpenOnload improves network buffer handling efficiency, enabling users to support higher message rates with predictable performance, ultra low latency, and bounded jitter. At such high message rates, competitive offerings can cause degraded system performance and deliver latency in the 10s if not 100s of microseconds with high jitter.

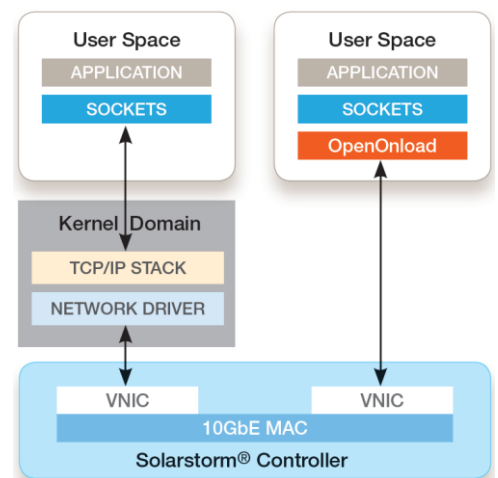


Figure 1: OpenOnload application

ETHERNET COMPATIBILITY

OpenOnload utilizes standard 10 GbE NICs and standard cost-effective Ethernet switches. OpenOnload operates over any Ethernet network including those in enterprise data centers as well as those in specialized environments, such as High Performance Computing (HPC) clusters, High Frequency Trading (HFT), storage networking, and other demanding applications. Furthermore, OpenOnload provides exceptional performance even if installed at only one end of a data link. In summary, OpenOnload is a flexible easy-to-deploy technology that integrates seamlessly into a wide range of enterprise data center applications.

TCP & UDP PROTOCOL COMPATIBILITY

OpenOnload significantly reduces host overheads and latency, while maintaining compatibility with the rest of the network. Since OpenOnload runs over standard Internet Protocol (IP), it allows standard IP routing, traffic engineering, and standard networking management and monitoring tools. OpenOnload is the most cost-effective method to lower latency and increase performance of TCP and UDP applications in 10 GbE networks, requiring absolutely no changes in the application or network.

POSIX SOCKETS APPLICATION COMPATIBILITY

OpenOnload can be used with a wide range of applications and integrates without the need for costly or time consuming application modifications. OpenOnload is fully compatible with the well-defined POSIX sockets API, so there are no changes needed to host applications. OpenOnload is binary compatible with applications, so it can run seamlessly in any application environment.

OpenOnload is a dynamically linked library that is simply invoked in the command line or environment with the application, so it does not involve configuring special transport protocols or NIC drivers, or modifying the application's network interface software. Since OpenOnload supports standard OSIX, it will accelerate applications over any Ethernet network – 10 GbE or GbE, as well as optical or copper cabling, and any combination.

RDMA OVER CONVERGED ETHERNET (ROCE)

RDMA over Converged Ethernet (RoCE) is also a kernel bypass technology. However, unlike OpenOnload, RoCE requires each application to be modified and standard Ethernet equipment to be replaced with specialized network components. This results from the fact that RoCE is an encapsulation of RDMA verbs (InfiniBand protocol) over Ethernet. RoCE requires that the Ethernet physical layer provide a reliable, congestion free transport. This requires a specially engineered version of Ethernet which is new and unproven, and can add to CapEx and OpEx costs.

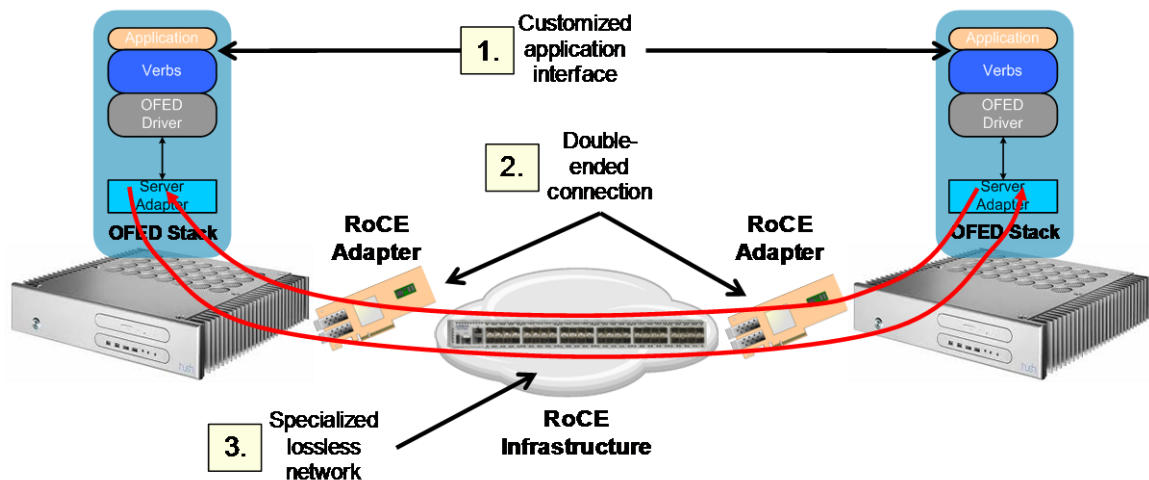


Figure 2: RoCE network implementation

Specifically, RoCE imposes a number of limitations and costs (some of which are illustrated in [Figure 2](#)) that can prove prohibitive:

Application customization: In order to achieve RoCE's performance benefits, applications must be modified to support RDMA verbs API. As a result, RoCE technology is inherently expensive to implement and limited in use cases.

End-to-end technology: RoCE requires both ends of the network connection, and the switching infrastructure in between, to be specially engineered to provide lossless data transfer, and can only operate between RoCE-enabled nodes. As a result, RoCE does not integrate well with the rest of the network infrastructure, creating a separately managed networking island.

Specialized network: RoCE cannot interconnect with a general purpose network in a broader data center or enterprise network environment, as it does not allow IP routing, firewall filtering, traffic engineering, nor standard networking management and monitoring tools.

Single-hop networking: RoCE networks are limited today to a single hop. Data Center Bridging Capabilities Exchange Protocol (DCBX - IEEE 802.1AB) is an emerging IEEE standard that will enable more complex multi-hop networks. However, DCBX has not been ratified and will take years to develop and mature. As a consequence, RoCE will be limited to single-hop networks until then.

Specialized protocol: RoCE uses a specialized network stack, including the InfiniBand transport protocol, that replaces the standard TCP/IP and UDP/IP network stacks. As a result, RoCE is incompatible with the most common forms of Ethernet traffic.

Additional CapEx: RoCE requires both ends of a connection and the switching infrastructure to be capable of supporting the new Data Center Bridging (DCB) technology. This often results in incremental capital expense for new equipment that would not otherwise be needed.

Additional OpEx: RoCE requires use of a complex software stack, typically the OpenFabrics Enterprise Distribution (OFED™), consisting of the RDMA verbs API, a replacement network stack, and kernel bypass services. Then the application software must be modified to support verbs, and the OFED software must be installed, recompiled, reconfigured, and re-tuned for each server host. This specialized network protocol requires specialized IT expertise in order to maintain and manage it. This increases operating costs, deployment time, and complexity to support a RoCE environment.

CONCLUSION

Solarflare's OpenOnload is a kernel bypass technology that is fully compatible with applications, TCP/UDP protocols, and standard Ethernet, making it the most flexible, easy-to-implement, low-latency solution on the market today.

Feature	Solarflare OpenOnload	RDMA over Converged Ethernet	Solarflare OpenOnload Benefit
Kernel bypass	✓	✓	Eliminates OS Overhead
Direct data placement	✓	✓	No copies – faster data transfer
Low latency / jitter	✓	✓	Critical for predictable performance
Ethernet compatibility	✓		Leverage existing network investment
Single-ended solution	✓		No network changes No end-to-end customized solution
TCP & UDP compatibility	✓		Accelerates commonly used protocols
POSIX sockets compatibility	✓		No application modifications No specialized network stacks & APIs
No incremental CapEx, OpEx	✓		Inexpensive to acquire, easy to deploy

Table 1: Feature comparison