

Requirements Document

Colorado1: Networking Requirements

Honsing Cheng
Sun Microsystems Inc.

Revision History

<i>Version</i>	<i>Comments</i>	<i>Date</i>	<i>Author</i>
1.1	Initial Draft	08/07/2008	Honsing Cheng

1 Introduction

SC has traditionally required the use of dedicated private networking hardware, including NICs and switches, to achieve a high level of reliability and better fault isolation.

- Dedicated hardware allows for some form of bandwidth guarantee and fault isolation at the link layer. Any overload or failure in the public network would not “spill over” and affect critical internal communications between SC nodes.
- SC can directly utilize, manipulate and manage the private network without interference from other applications. For instance, SC removes all unnecessary modules from all private networking code paths and inserts its own to send heartbeats in interrupt context.
- SC can also create private subnets for inter-node communication without involving the router or external network administration.

To further improve reliability, traditional SC has also required that two separate and disjoint dedicated networks be used. This has since been relaxed, as it proves too costly to implement for some deployments.

Colorado-I continues the trend of removing deployment barriers to SC. This document gives some details on the networking changes.

2 Goals and Non-Goals

This project seeks to remove the requirement of dedicated networking hardware for the SC private networks. These include NICs and network switches. In addition, no VLAN configuration is required on the switches to isolate private traffic. The user might still wish to do so and such config is supported. With this project, systems that are already connected to the same network (LAN, switch fabric, leased lines) can be made into a cluster without extra network hardware or switch config.

This project also includes any compatibility work required to support Project Clearview, which changes the IPMP architecture. It only affects public networking.

The following list highlights some existing behavior and restriction of the interconnect that are not being addressed by this project.

- All cluster nodes must be on the same layer-2 network, i.e. LAN, switch fabric, leased lines. That is, cluster nodes must be able to communicate without the use of a router.
- Internal cluster communication and the use of clprivnet (application traffic over the interconnect) do not support IPv6.
- The user must provide a IP subnet prefix for internal use by SC.
- Heartbeats are raw Ethernet frames with SAP=0x833. All other cluster communication, including CMM, is TCP/IP-based.

3 Requirements

3.1 Public Network

Project Clearview in OpenSolaris introduces a new architecture for IPMP based on pseudo-devices. SC relies heavily on IPMP for NIC monitoring and address failover.

- R1: SC shall be fully integrated with Clearview for the Colorado-I release. This includes changes to libpnm for functions related to the creation, tracking, and query of IPMP groups.

3.2 Private Network

Project Crossbow in OpenSolaris allows the creation of virtual NICs (VNICs) for the purpose of network virtualization. VNICs share the same hardware NIC, but each with different resources allocation, including on-NIC memory, TX/RX rings. Packets arriving at the hardware NIC is processed according to priorities and bandwidth control assigned to the VNICs. VNICs are presented as separate network devices to the rest of the system.

VNICs can be created without extra hardware or configuration on the network switches. They can be created even on older NICs, though features such as resource control might not be available.

Colorado-I can utilize VNICs to avoid the need for dedicated private networks while creating an abstraction of having dedicated NICs to the rest of SC. This helps reduce the development effort substantially.

- R2: SC shall be functionally compatible with the use of VNICs as its private network. This include heartbeats, internal communications, and application use of clprivnet.
- R3: SC shall support the use of VNICs that share the same physical NIC as the public

network as the interconnect.

- R4: SC shall provide an option to create VNICs during initial configuration. The user shall be given a choice of which hardware NICs to use and bandwidth allocation.
- R5: SC shall accept VNICs created by the user as private network adapter if they meet certain requirements, such as being on the same network.
- R6: The use of VNICs in Colorado-I is optional. Current support of dedicated NICs is continued.
- R7: Mixing VNICs and dedicated NICs shall be supported. However, when there is a large difference between the bandwidth of these NICs, performance could become unpredictable and may be limited by the lower speed NIC at peak load. Such configuration shall be discouraged.
- R8: NICs not supported by Crossbow shall be supported only as dedicated private network adapters for SC.

3.3 Security Consideration

The flexibility of sharing cluster private network and public network over the same physical NICs, cables, and switches comes at the cost of potentially less security on the private network. Malicious and ill-behaved machines can more easily gain access to the network where cluster heartbeats and internal communication are using.

- R9: Tools shall be provided to enable IPsec for both authentication and encryption of all cluster internal communications, including remote invocations, membership, and user data. Heartbeat packets however are not IP-based and are not covered by IPsec. They do not contain sensitive user data and can go in plaintext.

Despite the use of IPsec, several vulnerabilities still exist.

- 1) A malicious machine can spoof heartbeats of a down cluster member and mislead the surviving member into thinking its peer is alive, thereby preventing a necessary failover of services. The attacker must be on the same layer-2 network.

The spoof attack can be effective only for a short time (around 90 seconds). SC maintains TCP level connection checks for ACKs on the interconnect. If no ACK is received for more than 90 seconds, all connections to the peer are torn down and the peer is considered down, which in turn triggers a membership recalculation. Since the attacker cannot spoof TCP ACKs thanks to IPsec, SC will initiate the failover after 90 seconds.

- 2) A malicious machine can generate a packet storm on the public network using the public IP addresses of the cluster nodes. This attack can be launched anywhere on the public network. Routers to the cluster unknowingly forward the storm over the single physical NIC of the cluster node. While the node might not suffer from interrupt overload thanks to the polling architecture of VNICs, the switch port bandwidth might be overloaded enough that heartbeats cannot flow anymore. The cluster gets a split-brain.

The network must deploy effective DOS measures to safeguard both SC and non-SC

network services. SC can only rely on such measures when its interconnect shares the same hardware fabric. Being at the receiver end of a pipe, it has no way of preventing the pipe from getting clogged from the outside.

- 3) A variant of (2) is to launch the attack from the same layer-2 network as the cluster, which bypasses any IP-based DOS measures implemented at the router.

In one case, the attacker must gain physical access to the network. So physical access control measures might be sufficient.

In other cases, a machine already on the network must be hijacked and turned into a launching pad. Traditional anti-virus and anti-worm measures must be deployed on all systems of the network to prevent such attacks, both for SC and other network services.

4 References

- Project Crossbow: <http://opensolaris.org/os/project/crossbow/>
- Project Clearview: <http://opensolaris.org/os/project/clearview/>