White Paper | June 2009

radisys®

Achieving Backplane Redundancy in AdvancedTCA Systems

An Approach with Commonly Understood Standard VLAN Protocols on the Radisys Promentum® Platform

Overview

The Heart of “Five Nines:”System Redundancy

As customer expectations of communications service quality continue to rise, service providers (SPs) are driving toward “five nines” (99.999%) network uptime–which equates to approximately five minutes of unscheduled downtime per year. Network Equipment Manufacturers (NEMs) play an essential role in enabling SPs to achieve this ambitious goal. A well-designed, highly available network element not only reduces the chance of network failure; it also mitigates the downside of operational errors caused by external influences such as human error and natural disasters.

One of the key attributes of a highly available (HA) system is redundancy. This paper illustrates the concepts related to backplane redundancy on hardware platforms based on the Advanced Telecommunications Computing Architecture (AdvancedTCA or ATCA). These platforms are the foundation of a significant portion of the network elements powering today’s global wireless and wireline networks.

CONTENTS

Exploring Redundancy: Benefits and Trade-Offs

Background

ATCA is a PICMG[1] standard that addresses the needs of today's telecommunications network elements through the use of commercial off-the-shelf (COTS) products. The architecture defined by the ATCA standard is inherently fault tolerant in that power, shelf management and backplane interconnect are redundant.

For example, a Dual Star, Ethernet fabric ATCA system is by far the most common implementation of the ATCA standard, and is used in all of the scenarios in this white paper.

Figure 1 introduces three types of interconnect between hub/switches and node blades:

- *IPMI Management:* The I2C bus that is used for shelf management purposes like module hot swap, e-keying and power management.
- *Fabric Interconnect:* The interconnect network most commonly used for payload or data-plane traffic; it is usually Ethernet but is not limited by the standard.
- *Base Interconnect:* The interconnect most commonly used for management plane traffic. For low bandwidth systems, control and data plane could optionally flow over this interface instead of the fabric interconnect.

For certain architectures, base and fabric interconnects allow for the separation of management traffic from data plane traffic. This helps to avoid security breaches and improves session management. Each of the redundancy models described below can be independently applied to each path.

Redundancy Models: Options and Choices

Multiple redundancy models can be used to design a network element. N+M redundant models are applied from an application (services) perspective to design network elements that enable uninterrupted services. Most commonly, network elements employ two types of redundancy schemes: 2N and N+1.

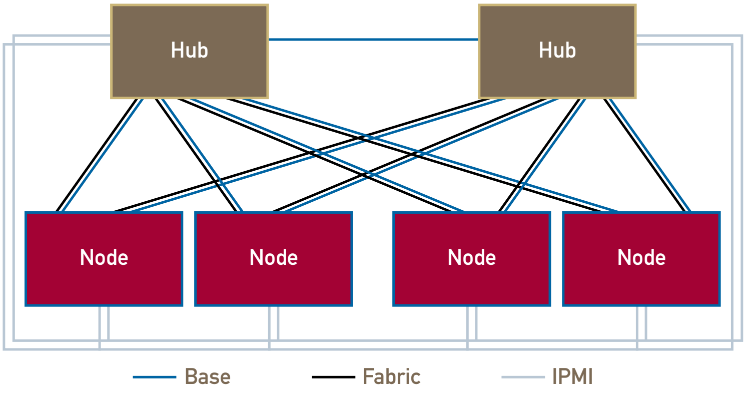

Figure 1. *ATCA Dual Star Topology*

- 2N simply means each resource has exactly one standby. This model allows for a hot standby that synchronizes state information from active, and takes an active role during a failover process. A 2N approach is typically used in management plane and control plane applications where state information cannot be lost. While a simple and effective basic method for achieving redundancy, 2N is cost-prohibitive; it doubles the hardware cost compared to a non-redundant blade.
- N+1 is much more cost-effective redundancy scheme. Here, the number of required resources is computed and one spare is added to act as standby. For instance, a typical 14- slot ATCA chassis may have 12 compute nodes, with 11 of them providing mission-mode processing. The twelfth nodes protects the 11 other blades, at a small incremental hardware cost compared to 2N redundancy.

In defining "five nines," unscheduled downtime does not include scenarios such as rebooting the machine after an important patch upgrade, service disruption during important configuration changes or upgrading to new hardware. Unscheduled downtime does include unexpected events caused by hardware, software or human failure. In manned equipment offices, the Mean Time to Repair (MTTR) is usually assumed to be 30 minutes. Assuming these numbers, a system must deliver uninterrupted function for six years or more.

The Critical Role of Middleware and Applications

Middleware software in high-availability systems manages the resources within the system to deliver important capabilities such as redundancy, fault detection, isolation and others. As a result, ATCA systems should be architected to keep such functions separate from the application, and also from the underlying operating system (OS) and hardware platform.

Specifically, middleware performs following roles:

- Continually communicates with system resources to collect operational data.
- Maintains awareness of how resources are mapped into redundancy groups.
- Checkpoints resources and adds or removes them from their respective group.
- Provides fault detection and isolation.
- Takes steps toward recovery.

Middleware generally is configured in 2N mode and provides functionality to realize redundancy at the chassis level, as well as at a system level when there are multiple chassis. It performs link and node fault detection and provides failover mechanisms.

Middleware software can optionally reside within the ATCA chassis. In systems utilizing the Radisys Promentum® ATCA-2210 10G switch, this software could be hosted on an embedded COM-E site or CPU blades that occupy two of the node slots. Typically these processors run the middleware's central control function. They can also perform functions like image storage and management–however, this topic is beyond the scope of this discussion on fabric redundancy.

Applications' Role in System Redundancy

In ATCA network elements, applications will dictate system performance requirements. For instance, some systems can tolerate a failover time as high as 500 milliseconds to one second. This failover window is not always adequate; specific application requirements often steer the failover scheme. Achieving faster failover usually involves more complex detection and management software.

Sometimes application software is designed to communicate down to the middleware when it detects potential problems in processing the payload. This information could be used by the middleware to mark resources as dysfunctional and remove them from service.

Redundancy Options in AdvancedTCA SystemsATCA System

ATCA systems can be configured in multiple ways to realize redundancy, based on individual customer requirements. These approaches include multicasts, spanning tree, Link Aggregation (LAG) and virtual local area networks (VLANs). As noted, different redundancy schemes will yield different system level configurations and behaviors during failover of the fabric switch or node blades. The paragraphs below discuss an illustration of switch element redundancy on fabric interface in single chassis and multi-chassis environments. The same concepts can be applied to base interface independently.

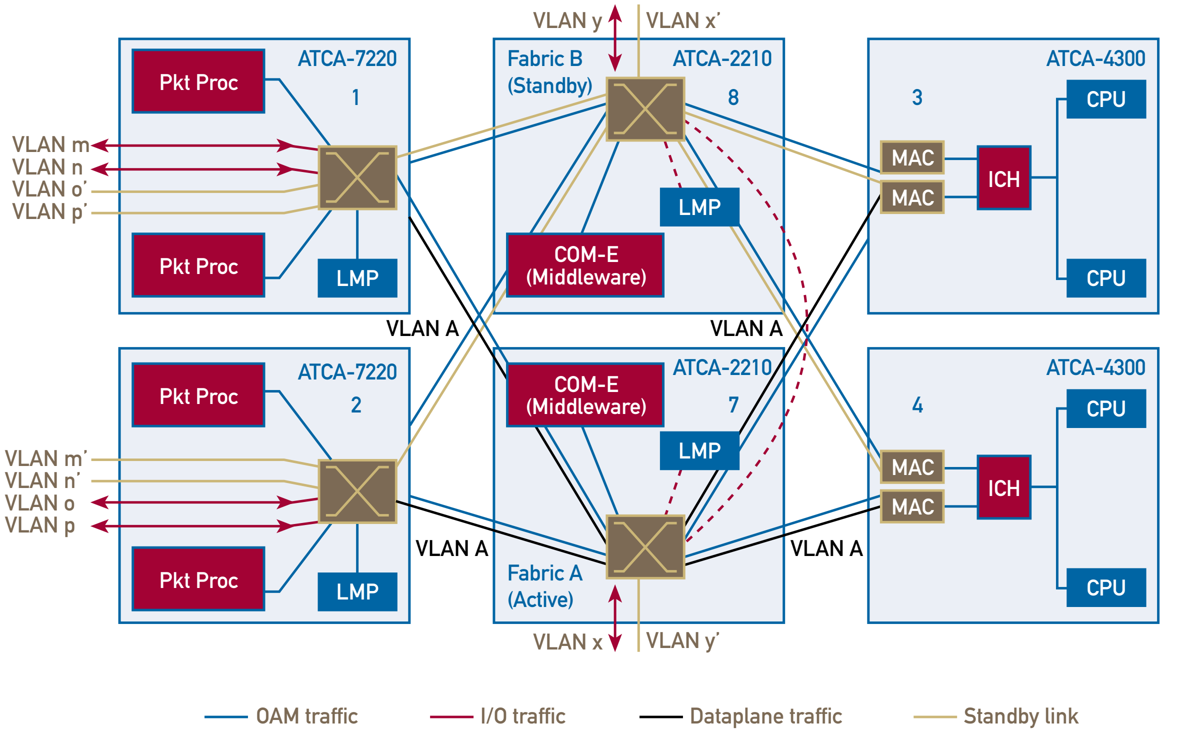

Figure 2. *Data Flow Model in an ATCA Chassis*

Figure 2 depicts a simple redundant network element based on an ATCA system with a dual star Ethernet backplane.

In this example system, there are two central switch hubs. They operate independently regardless of their role –e.g., both switches are functioning and able to forward packets at any time. Though the switches are independent, redundancy mechanisms on switch elements can be applied, depending upon the packet content and links used to receive and transmit traffic. Here, packet types can be divided into three broad categories:

1. Data plane traffic passes between switch and node blades (as well as inter-chassis links in multi-chassis architectures). In the figure above, the active/standby configuration on the switch fabric suggests that data plane traffic passes through VLAN A and can pass only through an active switch, i.e. Fabric A. All node blades transmit/receive packets in VLAN A on the backplane. All ports on Fabric A are dynamically added onto VLAN A. Similarly all the ports on Fabric B (except the single port connecting to Fabric A) are deleted from VLAN A.
2. Network I/O traffic passes between the external network and front/rear panel of ATCA blades. External traffic can directly enter into the system via a line card such as the Radisys Promentum® ATCA-7220 and/or ATCA-2210, but this traffic is considered data plane traffic when these packets are passed through the backplane. I/O redundancy can be optionally coupled with data plane redundancy as part of the system architecture. In the example above I/O traffic redundancy is decoupled with data plane redundancy. Fabric B is an active switch to receive VLAN y packets, routing these packets to Fabric A and similarly Fabric A is the active switch to receive VLAN x packets. Hence, both ATCA-2210s are in active-standby mode for I/O traffic with respect to certain VLAN packets. Also, since internal network configuration should not be exposed out on the external network, it is recommended that I/O traffic be handled using content-aware rules on ATCA-2210 and ATCA-7220 at layer3 to forward the packets onto the backplane.

3. Operations, Administration and Maintenance (OAM) traffic comprises management traffic flows between ATCA-2210 switch fabrics (and between ATCA switch fabrics and node blades hosted in the same chassis) to detect problems such as link, node and switch failures. Switches residing in the same chassis send and receive this traffic periodically to check the health of the system. A redundancy model is generally not applied to OAM traffic as this type of traffic flows through a combination of nodes and links to detect failures.

In this system, node blades like the Promentum ATCA-7220 and the Promentum ATCA-4300 will each have two fabric Ethernet links hardwired on the backplane to each fabric switch. The ATCA-4300 can use a bonding driver on its fabric links to manage these links and send VLAN packets over them. Applications communicate with the bonding driver to send packets on a certain link. On the other hand, ATCA-7220 can use port-based VLAN management on its switch to handle the packets on a certain link.

Failure Scenarios

Failures can happen on links, nodes in the system, external I/O or on the switch itself. Irrespective of the failure scenario, a system must be designed to handle each case with minimal traffic loss. Figure 3 shows a simplified view of a system to explain failure scenarios.

Node x is hosting an application with Node y acting as standby. Fabric A is the active switching element for data plane traffic and I/O packets on VLAN y. Wherein, Fabric B is a standby switching element for data plane traffic and active node for I/O packets on VLAN x. In this environment there are four common failure scenarios:

1. If the active Node x hosting application fails, the standby Node y becomes the active node for that application and continues services. To do so, the middleware software initiates a failover mechanism on the node blades.

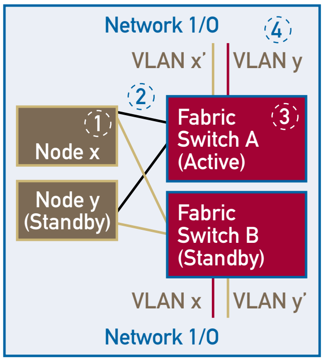

Figure 3. *Failover Scenario*

2. If the active link between the switch and node-passing data plane software fails, traffic is sent on standby link through Fabric B. Here, middleware software reconfigures all the node blades in the system to send traffic through a link connected to Fabric B, configuring Fabric B to be the backplane switching element.

3. If Fabric A fails, the active Node x for the application detects the link that failed and sends traffic through the standby link on Switch B. To execute this scenario, the middleware should function exactly as above in scenario 2.

4. If a network I/O on Fabric A fails, a corresponding standby link on Fabric B becomes active and sends traffic to VLAN y. Here the middleware software must act only upon the ports on each switch fabric, rather than the whole system. Again, the role of network I/O redundancy can be coupled or decoupled from data plane redundancy as part of the system architecture. (Note: In case of Virtual Router Redundancy Protocol [VRRP] implementation, Fabric B becomes the master router for network I/O.)

For an additional layer of assurance, middleware can check for failure scenarios described above and automatically trigger remediation, using configuration tools.

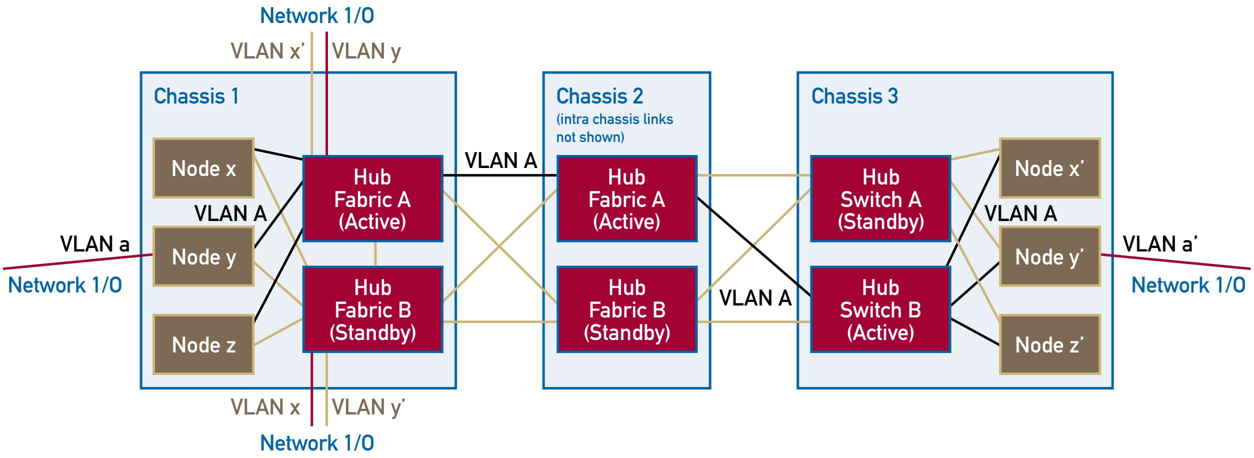

Figure 4. *Multi-Chassi Enviornment*

Achieving Redundancy in Multi-Chassis Architectures

With a goal of simplifying the network, redundancy among node blades can be spanned across multiple ATCA platforms, and the switches in each chassis can be in redundant mode to handle application redundancy. The same concepts that are used for single chassis, such as spanning tree and VLANs, can be extrapolated for multi-chassis environments.

In Figure 4, each of the traffic paths is assumed on separate VLANs. Network I/O traffic can enter from node blades or from switches in the various chassis, and are dealt by Layer 3 protocols for simplicity and security.

In this multi-chassis illustration, redundant node blades are spanned across Chassis 1 and 3, while Chassis 2 interconnects Chassis 1 and 3. Data plane traffic between different chassis is routed on a fixed VLAN A, and port memberships on a switching element are managed based upon the Active Fabric switching element on other chassis. For example, the port on Fabric A-Chassis1 connecting to Fabric A-Chassis 2 is part of VLAN A, wherein the port on Fabric A-Chassis1 connecting to Fabric B-Chassis2 is not.

Once the traffic comes into an active fabric, it is routed as the data plane path is followed to the next active switching fabric. Some system architects optionally configure multiple chassis in such a way that a failover of one switch in any chassis affects the Active/Standby role of switch pairs in other chassis, and changes the data plane route. In the example above, a system architect can force a switching failover on Chassis 1 when a failover occurs on Chassis 3 as part of requirements defined by service providers.

Conclusions

There are many ways to achieve redundancy in AdvancedTCA systems. The Radisys Promentum platform was developed to be flexible and accommodate a wide range of applications. ATCA-7220 supports Layer 2 protocols such as 802.11d (LAG) and spanning tree protocols that can be used to configure redundancy.The Promentum ATCA-4300 implements bonding drivers in a way that is similar to LAG implementation. The Promentum ATCA-2210 additionally supports Layer 3 protocols like VRRP that can support redundancy on network I/O.

In any network environment in which “five nines” system availability is the goal, the level of complexity to implement different redundancy schemes can vary greatly. The requirements of the application should drive the chosen approach. Across a wide range of applications, standard VLAN protocols can be used to achieve a robust redundancy on backplane. This will be sufficient to meet many requirements. To achieve faster failover, more complex–and in some cases, custom schemes–can be built by the application developer.

As the leader in providing ATCA based platform solutions, Radisys has the knowledge and experience to assist customers in architecting a solution that meets their requirements, while keeping development complexity at a minimum.

For more information about Radisys solutions and professional services for backplane redundancy, please visit www.radisys.com.

References:

[1] PCI Industrial Computer Manufacturers Group.

[2] The Radisys ATCA-2210 module supports both Layer 2 and Layer 3 networking protocols, but the word “switch” is used to keep the language simple.

radisys.

Corporate Headquarters

5435 NE Dawson Creek Drive
Hillsboro, OR 97124 USA
503-615-1100 | Fax 503-615-1121
Toll-Free: 800-950-0044
www.radisys.com | info@radisys.com

June 2009