Internet Draft
Expiration: June 2002
File: draft-ietf-gsmp-dyn-part-reqs-01.txt

T. Anderson
Intel Labs
C. Wang
Pacific Broadband Com.
J. Buerkle
Nortel Networks

December 2001

Requirements for the Dynamic Partitioning of Switching Elements

Status of this Memo

Conventions used in this document

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in [RFC2119].

Abstract

This document identifies a set of requirements for the mechanisms used to dynamically reallocate the resources of a switching element (e.g., an ATM switch) to its partitions. These requirements are particularly critical in the case of an operator creating a switch partition and then leasing control of that partition to a third party.

Definitions

In this document, the following definitions will be used.

Switching Element - A device that switches packets (e.g., an ATM switch or MPLS LSR) and whose resources can be divided into partitions, each of which can be independently controlled by a different controller.

Partition - A partition is a set of switching element (SE) resources. Partitions are also referred to as virtual SEs.

Active Partition - An active partition is a partition in which the resources are in use; either under the direct control of a separate controller or under internal policy based control.

Controller - The entity responsible for controlling the operations of an active partition.

Static Partitioning - In static partitioning, no changes can be made to any active partitionÆs resources without requiring a restart of that partition. Instances of repartitioning in which connections to controllers are disconnected before resources are reallocated therefore fall into this category.

Dynamic Partitioning - In dynamic partitioning, an active partitionÆs resources can be reapportioned without requiring a restart of the partition.

Frozen Partition - A frozen partition is an active partition that is in the process of being shutdown. A frozen partition's unused resources are relinquished, but all current connections are allowed to remain until removed by the controller. As connections close the resources are returned to the SE.

Deterministic Partitioning - In deterministic partitioning, each active partition is given an allotted quantity of each resource. The usage of resources in one active partition does not influence the resources available to another active partition. All discussions in these requirements presuppose the use of deterministic partitioning.

Statistical Partitioning - In statistical partitioning, some or all resources are pooled among the active partitions, and allocations may be based on percentages or on some other metric. Discussion of statistical partitions is outside the scope of these requirements.

Proactive Notification - A proactive notification is a message sent from a SE to its controller at the time an event occurs. Specifically, if a SE asynchronously sends the controller a message when it is dynamically partitioned, we say that the SE has proactively notified its controller of the resource reapportionment.

Explicit Reactive Notification - In explicit reactive notification, the SE does not asynchronously send a message when dynamic partitioning occurs. Instead, the SE includes a "resource changed" error code in the response to a subsequent request by the controller.

Implicit Reactive Notification - This is similar to an Explicit Reactive Notification except that the protocol does not contain an explicit "resource changed" error. In this case, all that the SE can do is to indicate that some unspecified error has occurred when the controller attempts to use non-allocated resources.

Introduction

Several logical entities are involved in the partitioning and control of a SE. First, a switching element (for the purposes of this draft) is a device that "switches" packets and whose resources can be partitioned and whose partitions can each be controlled by a single controller. (This partitioning also implies the ability to enforce this division of resources between competing partitions). Second, the partition manager (PM) is a management entity that specifies the number of virtual SEs into which the SE should be partitioned and the resources to be allocated to each virtual SE. Lastly, a controller directs the use of the resources of one or more partitions to provide a set of services.

In the rest of this draft, we will deal exclusively with logical entities although it is worth noting here that there are many possible mappings of logical entities to physical entities. For example, there may be multiple logical controllers running on a single physical processor (and for convenience we may refer to this processor as a physical controller). Likewise, there may be multiple partition managers running on a single management workstation. A switching element may consist of multiple physical elements (e.g., some number of blades in a chassis) or fractional physical elements (i.e., nested partitioning). Finally, any combination of these logical entities could theoretically be collocated on the same physical resources.

However, for many reasons, the physical realm often reflects this logical division of functionality. To facilitate this division, several protocols, such as MEGACO [RFC3015] and GSMP [GSMPv3], exist that allow control functionality to be physically separated from switching functionality. Recently, some regulatory environments have mandated multi-provider access to a single physical infrastructure. To satisfy these regulations, a common use of partitioning will be for the owner of the SE to partition the SE into several virtual SEs and then to lease these to third parties. In this case, the PM will likely be physically separate from all of the controllers. For locality (and therefore ease) of management, SEs will be remotely configurable and thus the PM will be physically separated from the SE. The following illustration depicts this arrangement. The dashed lines indicate interactions between the entities and are labeled with the cardinality of the relationship between the entities.

```
------------------             -------------------
|                | *         * |                 |
|    Partition   |-------------|   Controller    |
|     Manager    |      C      |                 |
------------------             -------------------
              1 \                / *
                 \              /
                  \ A        B /
                   \          /
                  * \        / *
                ------------/------
                |  --------/---   |
                |  |Partition |   |
                |  |          |   |
                |  ------------   |
                |Switching element|
                -------------------
```

Interaction A is one in which the PM partitions the SE and allocates resources to the partitions it creates. There is a one-to-many relationship between PMs and SEs. In order to support dynamic partitioning, this document will place certain requirements on proposed (or new) solutions in this space.

Interaction B is one in which the controller configures and manages an active partition. Current protocols implementing this interaction include GSMP [GSMPv3] and MEGACO [RFC3015]. These protocols allow a many-to-many relationship between controller and partition.

Interaction C is one by which a PM and a controller could communicate to alter the nature of an active partition. There is a many-to-many relationship between PMs and controllers. For example, there are multiple PMs per controller in the case where a controller is managing two partitions from different SEs and there are multiple controllers per PM in the case where a SE has two partitions each managed by a different controller. Possible types of interactions between PM and controller include:

- A controller could request that the resources of one of its active partitions be altered; either increased or decreased.
- The PM could respond to a controller request for altered resource levels.
- The PM could request that a controller release resources currently allocated to one of its active partitions. This could involve the following types of request:
  - A request to relinquish allocated but currently unused resources. That is to put a freeze on additional use of the specified resources.
  - A request to relinquish used resources.
  - A request to relinquish an active partition. That is a request that a controller release control of an active partition.
- The controllerÆs response to a PM request.

As far as the authors know, no proposed standard solutions currently exist for interactions of type C.

Dynamic Partitioning

Static repartitioning of a SE can be a costly and inefficient process. First, before static repartitioning can take place, all existing connections with controllers must be severed. When this happens, the SE will typically release all the state configured by the controller. Then, the virtual SE must be placed in the "down" state while the repartitioning takes place. Once the repartitioning is completed, the partitions are placed in the "up" state and the controllers are allowed to reconnect to the partitions. Then, the controllers can reestablish state in the active partition. Thus, static repartitioning results in a period of downtime and a period in which the controllers are reestablishing state. This is the case even if resources that are not currently in use in one partition, either an active or an inactive partition, are intended for a fully loaded active partition.

Therefore, dynamic partitioning is to be preferred to static partitioning since it avoids the downtime and loss of state associated with static partitioning. However, a different set of potential problems exists for dynamic partitioning. Some questions to be answered include the following:

- How is the controller notified of an increase or decrease in resources?
- What should happen when the PM would like to decrease the resources allocated to a partition but those resources are in

use?

Requirements

This document does not attempt to answer the preceding questions but instead defines a set of requirements that any solution to these problems MUST satisfy.

1. There MUST be a mechanism by which a PM can create virtual SEs on the SE and allocate SE resources to those virtual SEs.
2. SEs MUST ensure that controllers do not use more resources than those currently allocated to each virtual SE. Therefore, each control protocol MUST provide either an explicit reactive notification or an implicit reactive notification to indicate resource exhaustion.
3. Furthermore, this mechanism MUST support the partitioning of all resources discoverable through GSMP (e.g., label tables). Other resources used by GSMP indirectly (e.g., CPU) or resources (e.g., forwarding table entries) used by other types of SEs MAY be supported.
4. If a PM instructs a SE to release resources allocated to an active partition and if any of those resources are currently in use, the SE MUST deny the PMÆs request.
5. Subsequent to a resource reallocation failure, the PM SHOULD make use of one or both of the capabilities described in requirements 6 and 7.
6. A PM SHOULD be able to tell a SE to make an active partition into a frozen partition.

7. A PM SHOULD be able to contact the controller to ask it to reduce its resource utilization.
8. The PM MUST be able to exercise "power on/off" type control of the virtual SEs that it has created. When the virtual power to an active partition is turned off, the partition becomes inactive and any controllers associated with that partition are disconnected. This capability allows a PM to resort to static partitioning when a controller is uncooperative about releasing resources.
9. During dynamic repartitioning, a SE MUST maintain all existing state associated with the partitions being modified.
10. Control protocols SHOULD NOT include any mechanism by which a SE can ask its controller to reduce its resource usage.
11. Control protocols MAY contain proactive resource notification messages by which a SE could instantaneously inform the controller of an increase or decrease in resources. (We do not specifically require control protocols to contain proactive notifications because all control protocols must already have explicit or implicit reactive notifications as mentioned in requirement #2).
12. A PM MAY directly inform a controller of a change in virtual SE resources rather than rely on the implicit resource exhaustion mechanism of the control protocol.

13. SEs MAY inform the PM of resource exhaustion on a particular partition.
14. A controller MAY ask the PM for further resources or a reduction in existing resources.
15. To support the automation of interaction between the PM and attached controllers, the PM MUST be able to determine from the SE the addresses of the controllers that are currently attached to a virtual SE. Additionally, the SE MAY allow the PM to determine which control protocol (and version thereof) is currently managing each active partition.

Security Considerations

Only authorized PMs MUST be allowed to dynamically repartition a SE. Similarly, only the PM (or an authorized agent of the PM) that is authorized to partition a SE MUST be allowed to contact controllers to request that they decrease their resources or inform them that their resources have been increased. Likewise, the PM MUST verify and authenticate that any requests for additional/fewer resources for a virtual SE have come from a controller authorized to control the specified virtual SE.

Intellectual Property Considerations

The IETF is being notified of intellectual property rights claimed in regard to some or all of the specification contained in this document. For more information, consult the online list of claimed rights.

Acknowledgements

The authors would like to acknowledge the contributions of Avri Doria to this draft.

Normative References

[GSMPv3]   A. Doria, et. al, "Draft-ietf-gsmp-10.txt", work in progress.

[RFC2119] S. Bradner, "Key words for use in RFCs to Indicate Requirement Levels", RFC 2119, BCP 14, March 1997.

Informative References

[RFC3015] F. Cuervo, et. al., "Megaco Protocol 1.0," RFC3015, November 2000.

Author Information

Todd A. Anderson

Intel
2111 SE 25th Avenue
Hillsboro, OR 97124 USA
Phone: +1 503 712 1760
Email: todd.a.anderson@intel.com

Chao-Chun Wang
Pacific Broadband Communications
3103 N. First Street
San Jose, CA 95134
Phone: +1 408 468 6137
Email: ccwang@pbc.com

Joachim Buerkle
Nortel Networks Germany GmbH & Co. KG
Hahnstrasse 37-39
60528 Frankfurt
Phone: ++49 (0)69 6697 3281
Email: joachim.buerkle@nortelnetworks.com