# The Flooding Mechanism of the PNNI Routing Protocol

*Performance Aspects*

P. Jocher[1], L. Burgstahler[2], and N. Mersch[3]

1. *LKN, Technische Universität München, Arcisstr. 21, D-80290 München, Germany*
   *E-mail: jocher@ei.tum.de*
2. *IND, Universität Stuttgart, Pfaffenwaldring 47, D-70569 Stuttgart, Germany*
   *E-mail: burgstahler@ind.uni-stuttgart.de*
3. *Siemens AG, Hofmannstr. 51, D-81359 München, Germany*
   *E-mail: norbert.mersch@icn.siemens.de*

**Abstract** The Private Network Node Interface (PNNI) provides a flexible and scaleable routing architecture for ATM networks comprising a routing protocol and a signaling protocol. To obtain more experiences about PNNI, we developed a PNNI Emulator. We investigated the simple flooding mechanism of the PNNI routing protocol used to distribute topology information through the network. Beside theoretical results, the paper also presents some measurements of example networks.

**Keywords**: ATM, emulation, flooding, performance, PNNI, routing

## 1. INTRODUCTION

The *Private Network Node Interface* (PNNI), standardized by the ATM Forum (see [1]), provides a flexible and scaleable routing architecture for ATM networks comprising a routing protocol and a signaling protocol. PNNI routing includes mechanisms for the autonomous exchange of aggregated topology information to form a hierarchical representation of the network. Moreover, Quality of Service parameters are supported as required by ATM. PNNI signaling is based on a subset of UNI 4.0 signaling.

Our earlier investigations (see [2] and [3]) on principle aspects of the PNNI performance showed, that one of the crucial points is the load from topology information packet processing.

This paper is focusing on those performance aspects of the PNNI routing protocol concerning the flooding mechanism used to distribute topology information through the network. First, we look at the *PNNI Topology State Elements* (PTSEs) before we investigate the simple flooding mechanism. Besides theoretical results, this paper also presents measurements in example networks using our emulation tool developed at the Institute of Communication Networks/TUM in cooperation with Siemens AG.

The remaining part of the paper is organized as follows: Section 2 describes the basic characteristics of the distribution of topology information within PNNI. Section 3 concerns with the performance aspects of the PNNI routing protocol focusing on the flooding mechanism. Finally, section 4 concludes the paper and gives an outlook on future work.

# 2. PNNI TOPOLOGY INFORMATION DISTRIBUTION

## 2.1 Overview

PNNI uses source routing to determine a path through a network. Hence, every node needs a complete description of the topology to perform the necessary computations. However, when first turned on each node has only information about its own state. To complete a node's view of the network, the distribution of information must be provided by the routing protocol.

Section 2.2 describes the structure of the PNNI topology information groups; section 2.3 describes the distribution methods. These explanations are limited to a single peer group network.

## 2.2 Topology Information

The PNNI protocol provides a three leveled data structure for topology information. On the first level are the information groups (IG). Each IG only covers one specific part of a node, e.g. one port and its resources. On a second level, the IGs are bundled in PNNI Topology State Elements (PTSEs). Each PTSE contains IGs of only one type, so there are many different PTSEs describing each node.

PTSEs are the units of flooding and retransmission. As they do not contain information about the originating node, they need some kind of envelope when being sent to a neighbor. The PNNI Topology State Packet (PTSP) is such an envelope that transports PTSEs including information about the node's identity.

### 2.2.1 Information Groups

IGs can be divided into three classes:

- *Nodal information*: Nodal information includes the identity of a node, its capabilities, and information about the hierarchy. As long as there is no change in the hierarchy or no need to re-elect the peer group leader, the information in this group are static.
- *Topology state information*: Both, link and nodal state parameters, describing the characteristics of a link and a node respectively, belong to this group. Some topology state information are highly dynamic (e.g. the available bandwidth), while others are more static (e.g. the administrative weight). To keep a node’s topology information up-to-date, the dynamically changing IGs have to be distributed frequently.
- *Reachability information*: End-system addresses are contained in these IGs. Their number depends on the node’s role in the network. Nodes in access area may have many end systems attached (thus having many addresses in their databases), while nodes in the core network only might have a small number of attached end systems. As long as there is no mobility involved, their content is rather static and does not need to be distributed very often.

### 2.2.2 PTSE

PTSEs are used to bundle different IGs covering a certain aspect of a topology. While IGs only carry values that describe this topology aspect, the PTSEs also contain administrative information like a remaining lifetime or an IG identifier. PTSEs do not carry any information about the originating node.

Each PTSE can contain any number of IGs, provided they are of the same type and the PTSE does not exceed the maximum packet size. It is not necessary that all IGs of one type are bundled in one PTSEs. Rather, IGs can be bundled in a way that expresses a certain logical relation.

### 2.2.3 PTSP

To transmit PTSEs to a neighbor PTSPs are used. PTSPs contain at least one PTSEs of any type from a single originating node. Note, that only the PTSP reveals the source of the PTSEs in its header. For the receiving node, all information in a PTSP belongs to the same node. While it is recommended to transmit as many PTSEs in one PTSP as possible, the size of the PTSP must not exceed the maximum packet size.

## 2.3 Information Distribution

Two methods of distributing topology information are specified in PNNI: Database synchronization and flooding.

### 2.3.1 Database Synchronization

Database synchronization should happen rarely. Whenever two nodes learn for the first time that they belong to the same peer group, they exchange their complete database. They do this by announcing their database contents to the newly found neighbor. Then, the missing PTSEs are requested form the neighbor and finally exchanged.

### 2.3.2 Flooding

Flooding is a reliable method to distribute information within a network. Its main advantage, but also its main drawback is redundancy. On receipt of a PTSE that is not yet in its database, a node forwards this PTSE to *all neighbors*, except the one the PTSE was received from. If there is more than one path between any two nodes, a PTSE will be forwarded over each of them. Therefore, appropriate measures have to be taken, to prevent redundant PTSEs from consuming to much processing power at the receiving node.

Each received PTSE is checked, whether it is already installed in the nodal database. Following, there are two actions possible:

- Discarding the PTSE, if it is already installed in the database.
- Forwarding the PTSE via flooding, installing it in the database and then acknowledging of the PTSE.

There are two major reasons, why a node originally floods a PTSE:

- *Triggered Update*: Triggered flooding happens if a completely new PTSE is originated by a node or if there is a significant change in an IG within an existing PTSE (e.g. new end system addresses are added, the available bandwidth changed beyond a threshold etc.).
- *Aging*: Aging causes flooding if either the remaining lifetime of a PTSE reaches zero or if the remaining lifetime of the PTSE reached a certain threshold in its originating node. To prevent the PTSE from being deleted the originating node floods an update, even if the contents did not change.

Summarizing, while database synchronization is limited to the moment where two neighbors learn about their existence, flooding lasts as long as the network is up and running.

# 3. PERFORMANCE ASPECTS

In earlier investigations (see [2]) on PNNI performance we measured the processing load of a typical node[1]. Results showed, that with the given protocol stack about 80% of the load were due to PTSP processing while only 10% were caused by route computation. Moreover, in a hierarchical multi peer group network additional processing capacity is necessary for nodes representing their peer group at the higher network levels.

In the first section of this chapter we will give theoretical estimates of PTSE rates and their influence on the processing performance of a typical node. Following we are focusing on the simple flooding mechanism used to distribute topology information within the peer groups. Based on theoretical considerations and additional measurements we will show that - depending on the topology - a not to be neglected percentage of PTSEs is redundant.

## 3.1 Routing Protocol Processing

When a PNNI switch receives a new PTSE, it sends this PTSE to all neighbor nodes except the one the PTSE was received from. This is independent from the fact that some neighbors might already have flooded the same PTSE. It is also not altered by jittering the refresh interval, since jittering only influences the time a new PTSE is originated.

Therefore, we can follow, that if $adj_i$ is the number of adjacent nodes of node $i$, every PTSE originated by node $i$ must be

- sent: $adj_i$ times
- received: 0 times

and every PTSE not originated by node $i$ must be

- sent: $adj_i - 1$ times
- received: $adj_i - t$ times

$t$ refers to the fact that, due to the nature of flooding, a particular PTSE is not flooded upstream on those links, which form the shortest path tree (SPT) with the originating node as the root. Thus the value of $t$ depends on the considered originating node and receiving node. What follows is:

- The necessary routing protocol performance capacity of a PNNI switch increases linearly with the number of adjacent nodes, i.e. with the meshing of the network.
- Nodes, which own a number of PTSEs, which is above the average of a PNNI network, need less processing power for the flooding than nodes, with a smaller number of PTSEs.

[1] Square mesh topology with 25 nodes (single peer group) and 60% mean offered load.

This has to be taken into account when adding a node with low performance to a PNNI network. Based on this an estimation of the routing protocol processing capacity is possible and will be performed subsequently. However, we should take the following effects into account:

- Receiving PTSEs is more expensive than sending PTSEs. This is due to the big effort needed to decode the PTSEs and if necessary incorporate them into the database.
- The insertion of new PTSEs into the database is done only for the first new PTSE and not for the multiple duplicates received additionally.
- The processing time for a database insertion or check depends on the filling grade of the database.

It is clear that an estimation of the routing protocol processing capacity as performed below yields only a lower bound for the necessary performance capacity of a switch. Additional capacity must be provided due to the following reasons:

- PNNI timers are jittered but flooding is still bursty as our measurements in section 3.2.2 confirm. More capacity can prevent long queues.
- PTSE retransmissions due to bit errors must be taken into account.
- In case of the failure of a node or the insertion of a new node into the peer group database synchronization is required by the neighbor nodes of the failed or new node respectively.
- These considerations apply per peer group. Logical group nodes demand for additional capacity.

### 3.1.1 Flooding of Address PTSEs

This paragraph shows how the address PTSE processing performance of a PNNI switch depends on various parameters. The main factors influencing the address PTSE rate in a network are:

| | |
|---|---|
| – Number of address PTSEs in the network: | *NoAddrPTSE* |
| – Time between the refresh flooding of PTSEs: | *PTSERefreshInt* |
| – Network Meshing (average number of neighbors): | *NoNeighbors* |
| – The average value of *t* (in the network): | *T* |

Hence the average rate for received address PTSEs per node (*AddressPTSERate*) can be calculated as follows[2]:

$$AddressPTSERate \leq \frac{NoAddrPTSE \cdot (NoNeighbors - T)}{PTSERefreshInt}$$

[2] The '≤' relation refers to the fact that not every node necessarily originates address PTSEs.

Reserve capacity is needed, since the following parameters may change:

- The meshing of the network may be increased.
- The number of addresses within the network may increase with the consequence of an increased number of address PTSEs.

Within PNNI networks address summarization is applied to reduce the number of PTSEs to be flooded. Hence the address structure, i.e. the association of addresses to PNNI switches has a very big influence on the number of address PTSEs. Moving addresses within a network might deteriorate the summarization of addresses within peer groups, resulting in additional PTSE traffic.

### 3.1.2 Flooding of Link State PTSEs

Concerning the planning of the routing protocol processing capacity it is difficult to estimate the maximum rate of significant changes and thus the flooding rate, since this heavily depends on the dynamic behavior of the network and on the PNNI parameters. As an upper bound only the minimum interval between the flooding of PTSEs (*MinPTSEInt*) can be taken.

This results in the following formula for the average maximum rate of received link state PTSEs (*MaxLsPTSERate*) in a node within a PNNI network consisting of *l* links[3]:

$$MaxLsPTSERate < (2 \cdot l) \cdot \frac{NoNeighbors - T}{MinPTSEInt}$$

Extensive simulation is needed to evaluate the different influence factors of the update rate and to find a sensible upper bound (or increase the *MinPTSEInt*).

## 3.2 Flooding

### 3.2.1 Theoretical Results

We demonstrate in this section, that - depending on the topology - the simple flooding mechanism generates a not to be neglected percentage of redundant information in the network.

Figure 1 shows the flooding of an information element in an example network. Based on a significant event, *node a* originates a new information

[3] The ‘<’ relation refers to the fact that some of the link state PTSEs are originated by the nodes themselves.

element and forwards it to all neighboring peers (see figure 1-1). On receipt of this element, *node b* as well as *node d* checks if an instance is already in the respective database. If not, both nodes update their databases and forward the information to all neighbors, except the one it was received from (see figure 1-2). Thus, *node c* gets the same information twice. Because of sequential processing, one information element - in our example form *node b* - will be checked first, installed in the database and forwarded (see figure 1-3). Then, the second one will be processed and discarded, as an instance is already in the database. Thus, five information elements had to be processed in the example network. Only three of them would have been sufficient to update the databases of the respective nodes.

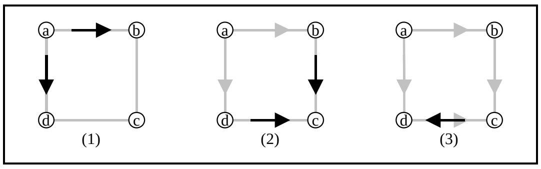

*Figure 1.* Example of the simple flooding mechanism

Due to the nature of flooding ‘overlapping’ may not occur on those links, which form the SPT with the originating node as the root. Hence, when a new information element is originated in a connected network consisting of *n* nodes and *l* bi-directional links, the following can be derived:

– Connectivity C:

$$C = \frac{2 \cdot l}{n}$$

– Number of information elements distributed on the SPT:

$$P_{spt} = n - 1$$

– Number of redundant information elements:

$$P_{redundant} = 2 \cdot (l - (n-1))$$

– Number of total distributed information elements:

$$P_{total} = P_{spt} + P_{redundant} = (n-1) + 2 \cdot (l - (n-1)) = 2 \cdot l - n + 1$$

– Ratio between $P_{total}$ and $P_{spt}$:

$$\frac{P_{total}}{P_{spt}} = \frac{(n-1) + 2 \cdot (l - (n-1))}{n-1} = 2 \cdot \frac{l}{n-1} - 1$$

The connectivity C is inappropriate to describe the behavior of a network with reference to the flooding mechanism, as table 1 shows.

*Table 1.* Properties of linear chains

| Topology | No. of nodes | No. of links | | C | $P_{redundant}$ |
|---|---|---|---|---|---|
| | | Total | SPT | | |
| Linear chain | 16 | 15 | 15 | **1.88** | **0** |
| Linear chain | 25 | 24 | 24 | **1.92** | **0** |

Therefore, we define a redundancy coefficient *R*, which implies information about the flooding behavior of the network:

$$R = \frac{l}{n-1}$$

The smallest possible value $R = 1$ only appears in connected networks consisting of a topology with a simple tree structure. Moreover, it follows:

$$\frac{P_{total}}{P_{spt}} = 2 \cdot R - 1$$

Table 2 shows the properties of some regular graphs. Already in the simple square mesh topology with 25 nodes 57% of the processed information elements are redundant. In the full mesh topology, this share increases to over 95%.

*Table 2.* Properties of regular graphs

| Topology | No. of nodes | No. of links | | R | $P_{total}$ | $P_{spt}$ | $P_{redundant}$ | $P_{total}$ / $P_{spt}$ |
|---|---|---|---|---|---|---|---|---|
| | | Total | SPT | | | | | |
| Linear chain | 25 | 24 | 24 | 1.00 | 24 | 24 | 0 | 1.00 |
| Ring | 25 | 25 | 24 | 1.04 | 26 | 24 | 2 | 1.08 |
| Square mesh | 25 | 40 | 24 | 1.67 | 56 | 24 | 32 | 2.33 |
| Full mesh | 25 | 300 | 24 | 12.50 | 576 | 24 | 552 | 24.00 |

Summarizing, because of the simple flooding mechanism, the number of redundant information elements caused by a single event strongly depends on the network topology. Consequently, paying attention to this fact in the network planning process would be one possibility to minimize the distribution of redundant information. Another approach is to modify the flooding mechanism. [8] proposes an interesting flooding method, delivering network updates faster than conventional mechanisms, while at the same time using significantly less bandwidth. However, an adaptation to our specific problem is necessary.

### 3.2.2 Measurements

To verify the theoretical results, we performed measurements on two emulated PNNI networks: A full mesh topology with 8 nodes and a square mesh topology with 16 nodes.

Both networks formed a single peer group with one end-system per node. Between any two neighbor nodes, there was only one bi-directional link. Each of them had a capacity of 155 Mbit/s (STM-1). The PNNI specific parameters had been set to values recommended in the annex of [1].

To simplify the measurements, we admitted only CBR (Constant Bit Rate) connections, though VBR (Variable Bit Rate) traffic could be easily supported by the use of equivalent bit rates (see [4], [5], [6], and [7]).

Bi-directional calls had been generated according to a Poisson process using three different call classes: 60% requested an 848 kbit/s connection, 30% requested a 4 Mbit/s connection and 10% a 12 Mbit/s connection. The link costs used for path computation were equal for all links. Hence, we varied the call arrival rate to adjust the mean offered load to 75%. The calls were equally distributed over the network according to a uniform random distribution of sources and destinations. The mean call holding time was 480 s. Table 3 contains additional properties of both networks.

*Table 3.* Network properties

| Topology | No. of nodes | No. of links | | R | $P_{total}$ | $P_{spt}$ | $P_{redundant}$ | $P_{total}$ / $P_{spt}$ |
|---|---|---|---|---|---|---|---|---|
| | | Total | SPT | | | | | |
| Full mesh | 8 | 28 | 7 | 4.00 | 49 | 7 | 42 | 7.00 |
| Square mesh | 16 | 24 | 15 | 1.60 | 33 | 15 | 18 | 2.20 |

Figure 2 shows the total number of measured PTSEs with a time granularity of 1 s in the full mesh network. Due to the small number of end-system addresses, most of the PTSEs are generated because of link state changes. This is also the reason why the refresh of address information is nearly invisible in figure 2.

One of the interesting characteristics is the plateau on the level of about 98 PTSEs/s and (in a weaker form) also on the level of about 196 PTSEs/s. Since all connections are bi-directional and a PTSE is generated for each direction of a link, every significant change causes $2 \cdot 49 = 98$ PTSEs (see table 3) to be processed in the example network. Additional measurements with reference to single PTSEs confirmed this.

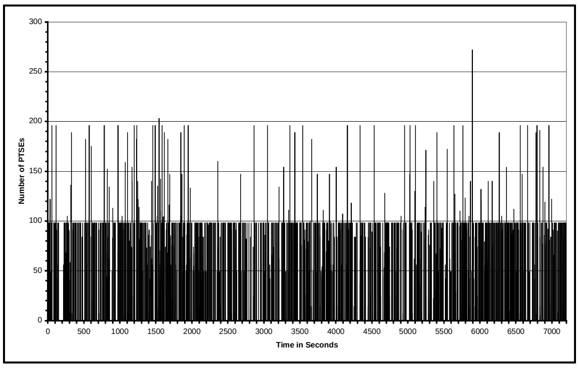

*Figure 2.* Total number of flooded PTSEs in the full mesh network

Consequently, the plateaus in figure 2 refer to significant change events. They occurred either on one link (98 PTSEs to be processed) or on two links simultaneously (196 PTSEs to be processed). In the latter case, it can be assumed that either two connections each spanning one link or one connection spanning at least two links has caused the significant change.

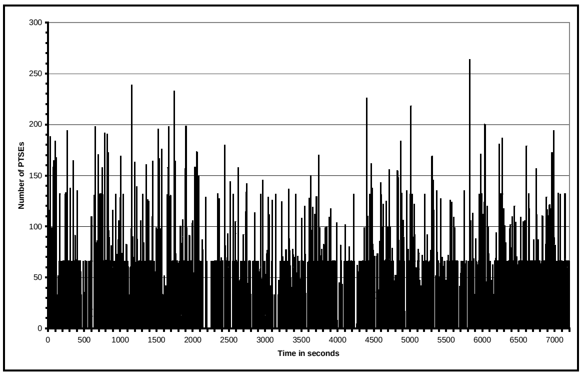

*Figure 3.* Total number of flooded PTSEs in the square mesh network

Figure 3 shows the number of measured PTSEs in the square mesh network. Again, a significant change generates a PTSE for each direction of a link. Thus, plateaus may occur at a multiple of $2 \cdot 33 = 66$ PTSEs/s (see table 3), as figure 3 confirms. The partly uneven formed plateaus are only caused by the time interval of 1 s used to measure the PTSEs.

Summarizing, our measurements confirmed the theoretical results. Additional investigations of several load scenarios have confirmed that the PTSE traffic is characterized by short peaks of high processing activity followed by periods of silence.

## 4. CONCLUSION

In this paper, performance aspects of the PNNI routing protocol have been presented focusing on the PNNI topology information elements and their flooding mechanism. The main factors determining the minimum routing protocol processing capacity of a PNNI switch are:

– the addresses structure (association of addresses to switches),
– the rate of significant changes on the links,
– the meshing of the network,
– the values of the various timers and thresholds defined in [1], Annex E.

The number of redundant PTSEs in a peer group caused by a single event strongly depends on the network topology. Consequently, paying attention to this fact in the network planning process is one possibility to minimize the distribution of redundant information. Another approach is the modification of the flooding mechanism. Therefore, further investigations are necessary and part of our future work.

## REFERENCES

[1] ATM Forum Technical Committee, Private Network-Network Interface Specification Version 1.0 (PNNI 1.0). ATM Forum af-pnni-0055.000, March 1996.

[2] U. Gremmelmaier, P. Jocher, J. Püschner and M. Winter, Performance Evaluation of the PNNI Routing Protocol using an Emulation Tool. 16th Int. Switching Symposium, Toronto, September 1997.

[3] P. Jocher, J. Frings, U. Gremmelmaier and M. Winter, Planning Aspects of ATM Networks using the PNNI Routing Protocol. NOC’98, Manchester, June 1998

[4] J. Roberts, U. Mocci, and J. Virtamo, Broadband Network Teletraffic. Springer, Berlin, 1996.

[5] International Telecommunication Union, Framework for traffic control and dimensioning in B-ISDN. ITU Recommendation E.735, May 1997.

[6] International Telecommunication Union, Methods for cell level traffic control in B-ISDN. ITU Recommendation E.736, May 1997.

[7] International Telecommunication Union, Dimensioning methods for B-ISDN. ITU Recommendation E.737, May 1997.

[8] Y. Huang, P. K. McKinley, Switch-Aided Flooding Operations in ATM Networks. IEEE INFOCOM `97, Kobe, April 1997.