

# Performance Study: Switched Networks vs Message Division Multiplexing

David M. Smith  
AbidaNet LLC

## Summary and Conclusions

This paper seeks to quantify the performance advantage of Message Division Multiplexing (MDM) over conventional switched networks by comparing its theoretical performance to measurements in simulation studies performed at Sandia National Laboratories. The goal of the Sandia study<sup>1</sup> was to determine the best switching network configuration to connect 256 general-purpose processors as a supercomputer cluster. In performing this study, they provided a set of baseline performance data against which we could compare network architectures based on MDM. There are two system performance parameters of interest: effective utilization of the available bandwidth and the inherent network latency in the message delivery.

There is no discernible difference between the four network topologies from the perspective of **bandwidth utilization**. All four are delivering data packets wrapped in similar packet preamble and header data<sup>2</sup> at a clock speed of 1 GBPS<sup>3</sup> with bandwidth utilization close to 100% of the theoretic availability for large message packets. They are also able to scale linearly with the basic clock frequency as the electronics and optics improve in performance.

Even a cursory glance at the charts below indicates a stunning improvement in **network latency** using MDM over the three switched networks. In three different tests in a pristine environment where the switches are expected to show an advantage, the MDM configuration at least holds its own, and in the case of broadcasting from a single source, out-performs even the Gigabit Ethernet configuration that also implements multicasting in hardware. In the tests that best simulate a real-world environment wherein messages are competing for the shared communications resources, the absolutely deterministic behavior of the MDM configuration completely out-performs the three switched networks, matched only in its minimum latency by one network. In the worst case scenario, it out-performs the best switched configuration by a factor greater than 6, and the other two with over 60 times less latency. In almost all cases, the worst latency performance is the Gigabit Ethernet, perhaps the most likely competitor for the MDM network in the commercial market place.

In addition to out-performing any switched system, MDM architectures offer the following advantages:

- ❑ Dropping inactive stations with less than 1% loss of bandwidth,
- ❑ Adding new stations with no significant interference with the existing data traffic with an upper limit of at least 10,000 stations on one network, and
- ❑ Built-in security that validates authorized stations before they connect, and can geographically localize unauthorized attempts at access.

---

<sup>1</sup> Chen, Helen and Wyckoff, Pete "Simulation studies of Gigabit Ethernet versus Myrinet using real application cores", CANPC'00 workshop of High-performance Computer Architecture, Toulouse, France, January 2000

<sup>2</sup> One specific MDM implementation needs a slightly larger header to encapsulate additional network dynamic and statistical information

<sup>3</sup> Giga-bits per second, bits per second \* 10<sup>9</sup>

When you add to this stunning performance the flexibility inherent in this approach, the only question remaining is whether network topologies based on MDM can be realized at a practical cost. Work is currently under weigh at AbidaNet LLC to develop the first prototype of such a system with a target cost of the prototype equivalent to the actual costs of the switched networks in this report.

## ***Background***

MDM is the heart of a revolutionary approach to networking. Instead of connecting stations by means of a network of switches and routers, all stations connect via bidirectional couplers to a Fiber Optic bus, which can contain an arbitrary number of branches. The classical approach to managing access to such a bus, the IEEE standard 802.3 Ethernet protocol, permits each station to launch an attempt to transmit whenever the data are available. If this attempt collides with other stations' attempts to transmit, each backs off a random amount of time and re-transmits. While this works tolerably well (although limiting the users to around 30% of the available bandwidth) at 10 MBPS<sup>4</sup> and 100 MBPS, when the carrier frequency reaches 1 GBPS or more, other well documented practical constraints make such a topology totally impractical. Networks using Ethernet at these speeds resort to connecting the individual stations by a dedicated link to a collection of routing switches that eventually deliver the data to its destination.

MDM takes a different approach with two key enabling characteristics:

- ❑ Every interface between a station and the optical bus has the ability to restore to the bus the optical energy tapped off to the interface, and
- ❑ Intelligence on the interface card organizes the stations into an optimal sequence in which they are permitted to transmit. The optimality minimizes the propagation delays across the network.

Every transmission onto the fiber is therefore guaranteed to be collision-free, and all stations have access to that transmitted data delayed only by the propagation time of the data on the fiber.

## ***The Simulation Study***

Helen Chen and Pete Wyckoff performed a very careful study to assess three (in spite of the title of their paper) different architectures for connecting 256 processors. In each case except as noted below, the processors were connected to the switching layer(s) by 1GBPS full duplex cables from standard, commercial network cards. The architectures were:

- ❑ Conventional Gigabit Ethernet switches cascaded because at that time, the switches were limited to 64 connections. The cost of the 5 switches was \$150,000.
- ❑ A Myrinet configuration using 32 16 port cross-bar switches for a total cost of \$160,000<sup>5</sup>, and
- ❑ An Avici Terabit Switch Router (TSR) that used twelve 20-GBPS full duplex links to form two 3-D toroidal meshes connecting all the station interfaces at a cost of \$250,000.

They ran a series of simulation tests on these configurations in each case measuring the time between a processor creating a message and the time of reception at the destination. The time difference was measured and accumulated statistically. One important parameter was omitted from the description: the mean distance between the processors. However, since they observed that the Myrinet switches were limited to a distance of 10m from the host, we presume this is an

---

<sup>4</sup> Mega-bits per second, bits per second \* 10<sup>6</sup>

<sup>5</sup> The Myrinet base frequency was actually 1.28 GBPS, and required special interface cards rather than the standard network cards. To establish a consistent performance baseline, we multiplied all Myrinet latencies by 1.28.

upper bound on the station separation, and performed some simple analysis to estimate the actual average separation between processors to be about a foot.

The tests performed were as follows:

1. **token\_pass** – passing a token around a virtual loop to measure both the minimum message latency by sending empty messages and the bandwidth utilization by sending large messages.
2. **fan\_in** – 255 processors simultaneously sending a message to the same destination
3. **fan\_out** – one processor broadcasting a message to 255 destinations
4. **mesh** and **torus** – two parallel algorithms running on all the processors arbitrarily exchanging data as necessary. While the **torus** algorithm was a little more I/O intensive than the **mesh** algorithm, Chen and Wyckoff reported no significant difference in network latency between the two algorithms.

The results from these studies were used as a reference point for comparing the latency of the MDM configuration. The only reference they make to the bandwidth results from the **token\_pass** study was “the simulation throughput values were different from the corresponding theoretical bandwidth by less than half a percent.” Even with the Ethernet protocol, this is not surprising since in that test, only one processor at a time was transmitting.

## ***MDM Analysis***

Analysis of an MDM network is straightforward. When a station is allowed to transmit, it places a 256 bit preamble, a 96 bit header, the payload and a 16 bit trailer onto the bus. We assume also a 10 bit “dark spot” between messages on the bus. This message propagates in the fiber at 5ns per meter to all the other stations.

## **Minimal Latency Analysis**

The **token\_pass** study with zero size payloads provided the simulation systems’ minimal latency. The equivalent MDM minimal latency is computed by transmitting the packet to its immediate neighboring station. In our case, this is an actual ring rather than a virtual ring. The minimum latency computed here will be the same for all stations. However, we need a little more analysis because we do not know the mean distance between stations. The MDM data were therefore generalized to compute the average and maximum latencies assuming that the virtual neighboring station was not the neighbor on the physical fiber. The worst case would have the virtual neighbor separated by 254 other stations from the source. The average case would separate them by 127 stations.

The unknown in the comparative simulation data is the mean distance between the stations. However, we can now estimate this from the results of the minimum latency analysis in Figure 1. Recalling that the network topologies connect all the stations to one or more switches, we could argue that the best latency they could achieve in this pristine environment would be some measure of the mean distance between stations. Consequently, as we compare the data in Figure 1, we see that the best latency performance is achieved by the Myrinet configuration, and this is roughly equivalent to the MDM data with 300 cm spacing between stations. We will therefore use this value in subsequent computations.

## **Fan-in Study**

The **fan\_in** study had all the stations simultaneously transmit a 2048 byte message to one destination. For MDM, the latencies for this exercise would be calculated as the sum of the time a station waits to transmit, the time to inject the message onto the bus and the propagation time to

the destination. The worst case is when a station has to wait 256 message times to transmit, and the destination station is at the farthest extreme of the bus. The best case is when the opportunity to transmit arrives immediately and the destination is only one station away. Since the MDM delivery is absolutely deterministic in this test, it is easy to argue that the average latency is the mean of the best and worst cases.

## Fan-out Study

In the **fan\_out** study, one station transmits a 2048 byte message to be received by the other 255 stations. For MDM, since broadcast is a hardware feature, this costs the time to transmit one message and then the minimum and maximum propagation times to the other stations.

## Mesh and Torus Studies

The **mesh** and **torus** studies attempted to replicate operational activities in the 256 node cluster. Messages of a specific size were generated at reasonable rates, but random intervals, and the latencies between source and destination measured. Since the MDM network architecture is completely deterministic, the analysis is once again very simple. In the best case, the message is delivered from one station to one of its two neighbors. The worst case was to deliver a message to a station 255 links away, and the average case 84 links away<sup>6</sup>. We make the same simplifying assumption made in the Chen/Wyckoff data that the processing at each station is synchronized to the I/O ports so that the network interface card is able to transmit as soon as the data packet is available.

## Results

Even a cursory glance at the charts below indicates a stunning improvement in network latency using MDM over the three switched networks. While the three switched networks all have moments (in different test situations) where their performance comes close to that of the MDM configuration, each also has situations where their performance is surprisingly poor. Even in the first three tests that emulate a pristine environment where the switches are expected to show an advantage, the MDM configuration at least holds its own, and in the case of broadcasting from a single source, out-performs even the Gigabit Ethernet configuration that also implements multicasting in hardware.

## Minimum latency

Figure 1 shows the minimum latency comparisons between three MDM configurations and the test configurations for the **token\_pass** test. In addition to demonstrating the baseline latency in the four candidate configurations, these results were used to estimate the mean distance between stations that was not provided in the reported results.

The MDM configurations used mean distances between stations of 100 cm, 300 cm and 1 m. Since the performance at 300 cm most closely the Myrinet results, we conclude that the mean distance between processors was about 300 cm, or roughly one foot. This would be consistent with processors rack-mounted in groups. As will be seen later, the Myrinet configuration best approximates the MDM performance under ideal circumstances. This is because there are inescapable latencies built into the Avici and Gigabit Ethernet architectures that are not present in Myrinet or MDM configurations.

---

<sup>6</sup> Since there are many more short hops than long, the average distance is 1/3 of the maximum, not 1/2 as one might expect.

Minimum Latency Comparison

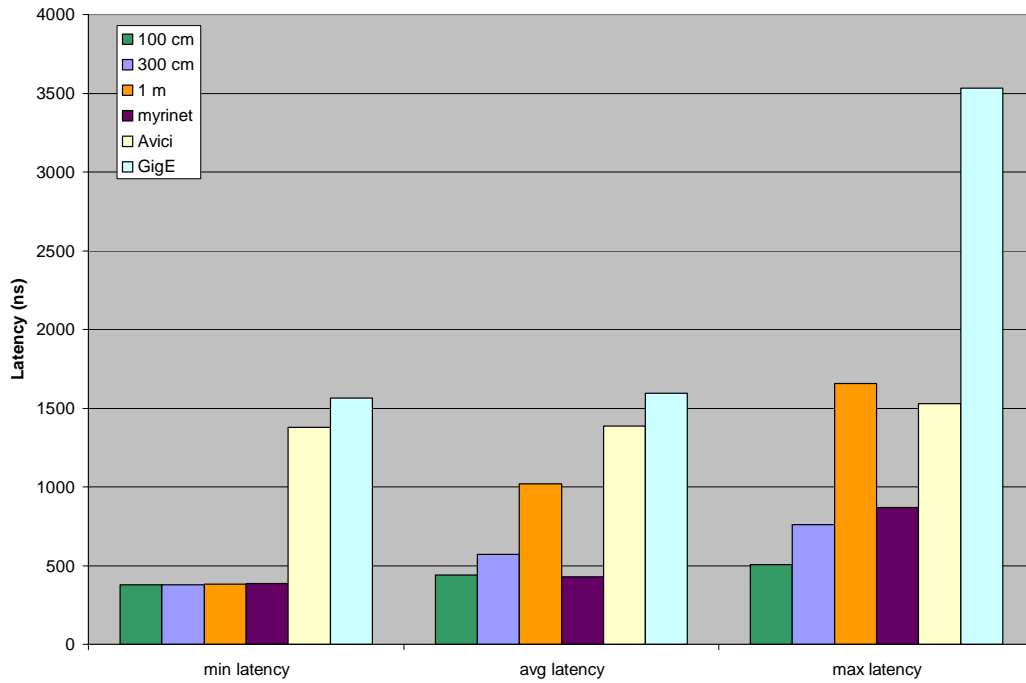
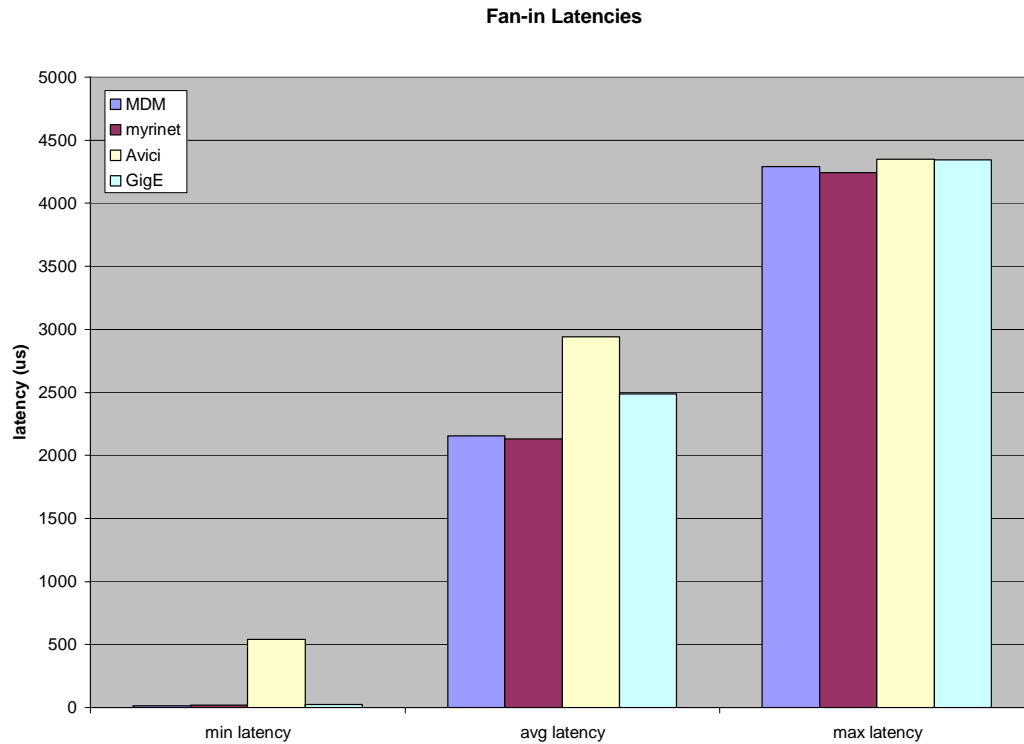


Figure 1 - Minimum Latency Comparison

### Fan-in Study

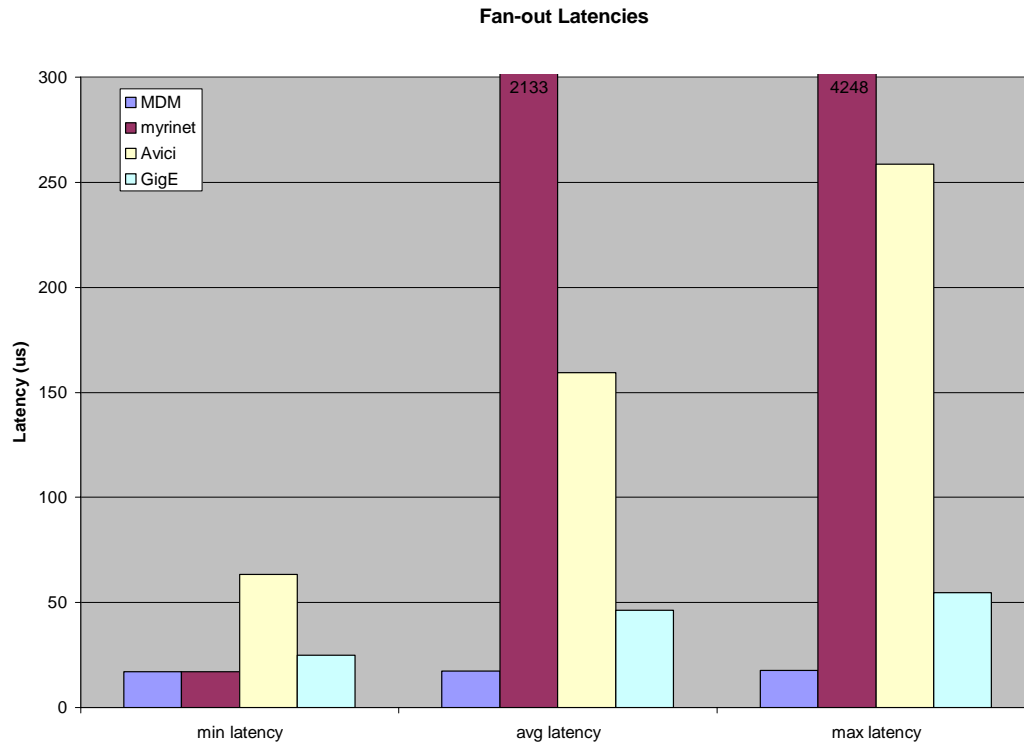
Figure 2 shows the fan-in comparative results where all the processors simultaneously attempt to send messages to a common destination in an otherwise clean data environment. Because of its essentially serial nature, the MDM configuration did not show any particular latency advantage over the switched networks that use the switches to ‘cut across the circle.’ Its high minimum latency is probably the worst performance for the Avici configuration that otherwise performed very well in the subject study.



**Figure 2 - Fan-in Study Results**

### Fan-out Study

Figure 3 shows the results of the fan-out study. Surprisingly enough, the only simulation configuration close to the MDM performance is the Gigabit Ethernet because, like MDM, the multicast facility is implemented in the hardware. While the Avici configuration was able to make some attempt at distributing a single message albeit with some latency, it appears that the Myrinet configuration forced the source to send multiple messages. Unfortunately, in real supercomputer topologies, it is not uncommon for one processor to need to broadcast state, status or commands to all the others, and this is a huge performance hit for the Myrinet configuration.



**Figure 3 - Fan-out Study Results**

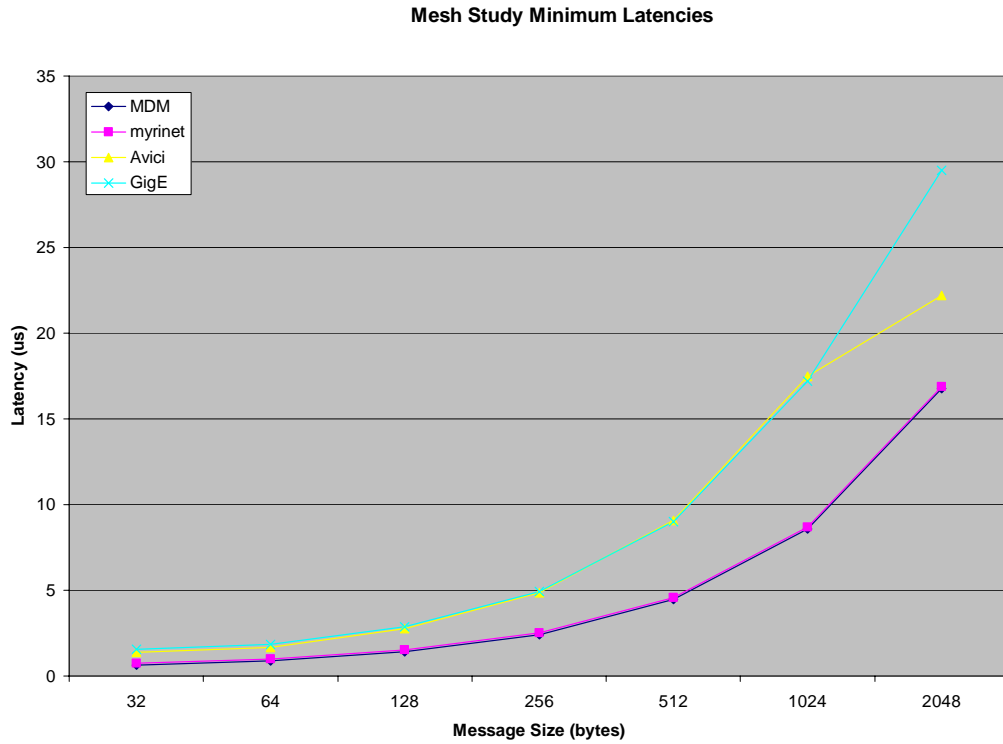
## Mesh Study Results

The following charts show the minimum, average and maximum latencies as a function of message size for the same four configurations in the **mesh** and **torus** studies. In these tests that best simulate a real-world environment wherein messages are competing for the shared communications resources, the absolutely deterministic behavior of the MDM configuration completely out-performs the three switched networks, matched only in its minimum latency by the Myrinet configuration. This match is actually a powerful validation of the assumptions underlying the emulation of the MDM configuration. In this particular optimal situation of no internal interference from other message paths, the message delivery latency is dominated by the time necessary to physically serialize the message packet onto the data bus. The Myrinet configuration uses only one switch path without contention to reach another processor in the best case. Even in this ideal scenario, there are latencies built into the switch environment both of the Avici and Gigabit Ethernet implementations as indicated in Figure 1.

In the worst case scenario, the MDM configuration out-performs the best switched configuration (Avici) by a factor of 6.5, and the other two by a factor of over 65. The conclusions noted in the Chen/Wickoff report attribute the dismal latency performance of Gigabit Ethernet and Myrinet configurations to contention and queueing problems in the layers of switches due to simultaneous message traffic between other pairs of processors. This same contention does not hurt the Avici configuration so badly because of the high-performance internal switches and busses.

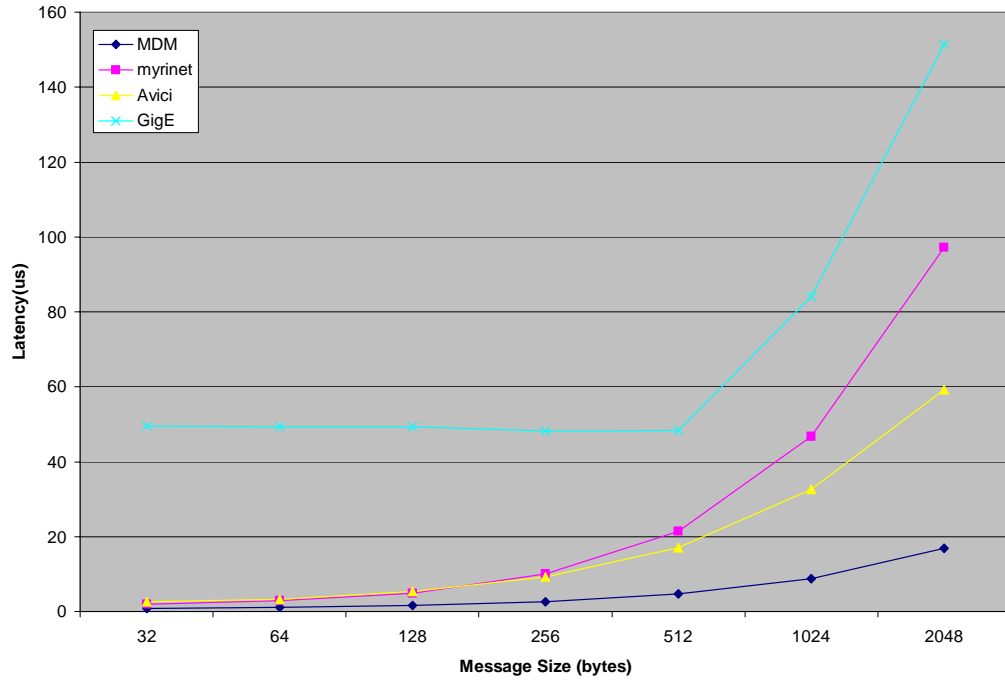
In all cases except the worst-case latency for large messages, the worst latency performance is the Gigabit Ethernet, perhaps the most likely competitor for the MDM network in the commercial market place. At the time of testing, this was aggravated by the fact that switches were not available with more than 64 ports. Consequently, to interconnect 256 processors, they had to use four 64 node switches and one second-stage switch node to deliver packets between them.

Contention for this second-stage switch becomes a serious bottleneck to traffic throughout the network unless two stations communicating happen to be on the same first-stage switch. This contention is most noticeable with small message sizes where the existence of any message in a queue regardless of its size is going to delay delivery by the same amount of time.



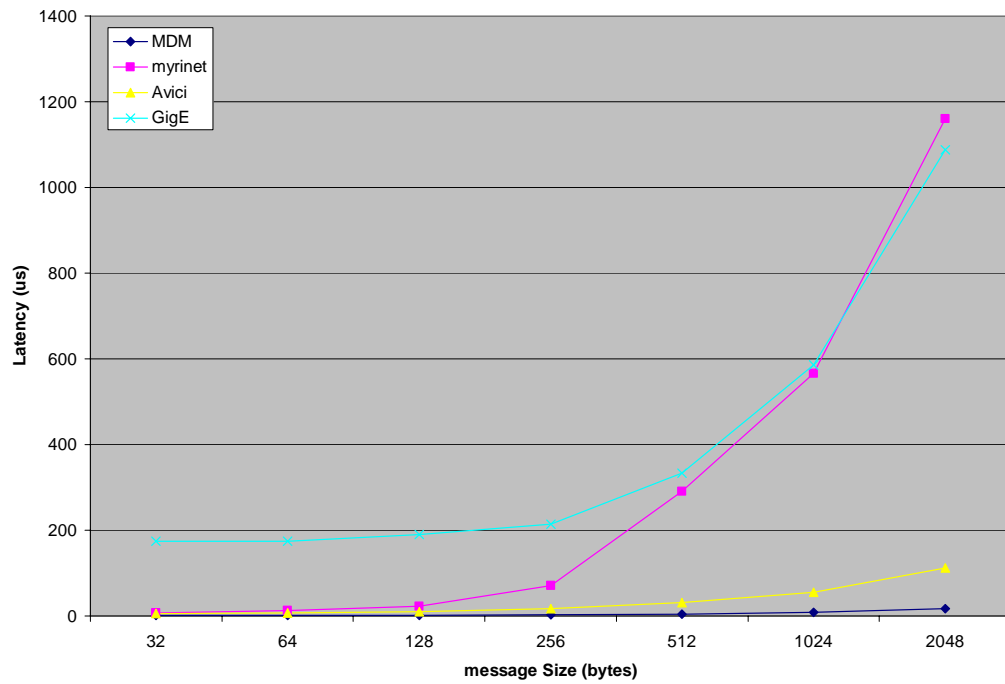
**Figure 4 - Mesh Study Minimum latencies**

**Mesh Study Average latencies**



**Figure 5 - Mesh Study Average latencies**

**Mesh Study Maximum latencies**



**Figure 6 - Mesh Study maximum Latencies**