ELSEVIER

Computer Networks and ISDN Systems 28 (1996) 583–588

COMPUTER NETWORKS and ISDN SYSTEMS

# Lessons learned from a MAN

Hank Nussbacher *

*Israeli InterUniversity Computer Centre, Tel Aviv University, Ramat Aviv, IS-69978 Tel Aviv, Israel*

**Abstract**

The Israeli network converted to using a Metropolitan Area Network back in February 1994. This MAN network provides coverage over 250 kilometers and 3 major cities in Israel. The seven Israeli universities use Ethernet ports to the MAN running at 10 Mbit/s. This paper describes the lessons learned from running such a nationwide Ethernet network.

## 1. Background

Many European organizations are exploring Ethernet port access as an economical method of accessing an ATM network. the ATM Forum is advancing the L-UNI standard (LAN Emulation User-to-Network Interface) which will make Ethernet connections to ATM a worldwide standard. This paper will hopefully explain to potential network administrators the advantages and disadvantages of using an Ethernet access method for their national connections. The network analyzed is not an ATM network but rather a precursor to ATM networks—the DQDB MAN network.

Israel has been running a nationwide Metropolitan Area Network (MAN) since February 1994. The topology that has been installed is as follows:

| | | |
|---|---|---|
| Weizmann Institute | February 1994 | weizmann.ac.il |
| Hebrew University | February 1994 | huji.ac.il |
| Bar-Ilan University | February 1994 | biu.ac.il |
| Tel-Aviv University | February 1994 | tau.ac.il |
| Technion | April 1994 | technion.ac.il |
| Haifa University | August 1994 | haifa.ac.il |
| Ben Gurion University | November 1994 | bgu.ac.il |

Our PTT, called Bezek, installed equipment from Alcatel Bell to implement a country-wide MAN. There are three PTT switches, located in Jerusalem, Haifa and Tel-Aviv. The geographic spread of our academic network covers an area of approximately 250 km. The MAN network is a DQDB (Double Queue Double Bus) network running on 34 Mbit/s circuits. Our PTT offered a

* E-mail: hank@vm.tau.ac.il.


*SSDI* 0169-7552(95)00087-9

service known as "Ethernet passthrough" which allows each university to connect via Ethernet to the others. What we have established is a nationwide Ethernet (10/Mbs) using available Ethernet ports on our cisco AGS/4 routers that currently (October 1994) moves over 300 gigabytes per month, see also Fig. 1.

## 2. Installation

Installation of the MAN network required almost no changes to our cisco routers. The connection appears as a normal Ethernet port with no special parameters to the cisco. We currently route IP, IPX, Appletalk and Decnet over the 10 Mbit/s MAN network.

## 3. Management headaches

One of the first things we learned from this system is that SNMP becomes close to worthless. When a leased line or Frame Relay circuit goes down, the cisco generates an SNMP trap that can be picked by various SNMP management programs and alert people and operators about the event. When the connection is lost to a particular site via the MAN, there is no trap created. It is

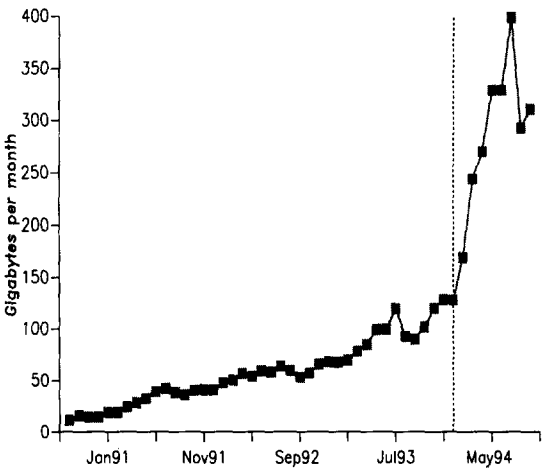


Fig. 1. Monthly traffic in Israel.

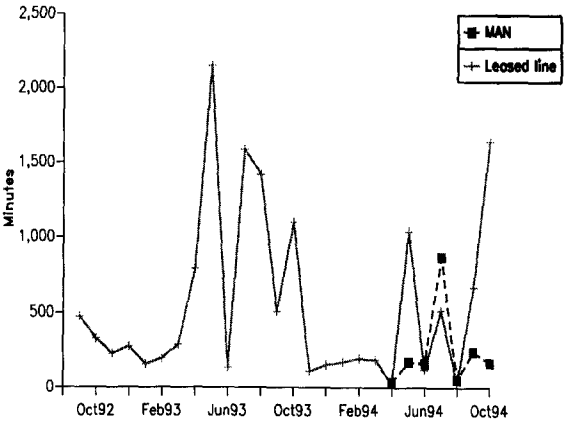


Fig. 2. Downtime in minutes per month.

the same as if a Unix system was unplugged from an Ethernet LAN.

The workaround we found was to create ping daemons that run in different places on the network and that ping each of the seven cisco Ethernet MAN ports (one per university), once every minute. If the ping doesn't succeed, the daemon immediately tries 2 more times before creating a pseudo-trap event and informing all concerned parties that the MAN to a particular university is down.

Next we found that we had placed all our eggs in a very sensitive basket. Previously we had used redundant digital leased lines in a partial mesh configuration. Whenever the line would go down, IGRP would just route around the problem. Every university had at least two outgoing leased lines to different neighbors and some had as many as five. The MAN was touted as 100% reliable, which we found to be not completely true, see Fig. 2. Whereas for leased lines we experienced 623 minutes of downtime per month on average, with the MAN we experienced 241 minutes of downtime per month. This was a significant reduction but the single point of failure meant that for 4 hours per month, each university was cut off from the Internet.

This was an intolerable situation so we started to install a Frame Relay backup network (256 kb circuits with CIR = 0) in October 1994. The PVCs are designed in a ring topology so that every

Table 1

| Line speed | From-to | Medium | 8000 octet ping (s) | 1000 octet ping (s) | 32 octet ping (s) |
|---|---|---|---|---|---|
| 10 Mb | Local LAN | Ethernet | 0.024 | 0.004 | 0.001 |
| 10 Mb | Israel–Israel | MAN | 0.048 | 0.020 | 0.004 |
| 100 Mb | Local LAN | Fibermux TDM | 0.048 | 0.012 | 0.004 |

university has two neighbors to which it has a logical circuit.

In addition the FR topology is designed so that if any one of the three MAN switches fails, all sites have a connection to at least one site that is not located on the same MAN switch.

The most serious problem we have had with the MAN network though is one where not only SNMP management breaks but also where ping daemons can't help. This is what we call "one way" transmission. We have had 3–4 cases where the systems connected to the MAN switch start behaving in a simplex fashion—they can either send or receive data—but not both.

This problem has proven to be difficult for our PTT and Alcatel Bell to track down and fix. We suspect it has to do with some sort of misconfiguration of the broadcast tables inside the switches or that the software running the switches gets confused when new nodes get added to a broadcast group. Unfortunatley, the MAN network appears to us as an Ethernet and we are unable to debug the PTT network for them.

## 4. Benchmarking

Benchmarking the MAN has proven to be a challenging task. We conducted many different benchmarks. First we looked a the raw numbers that a DQDB network can sustain. Every DQDB segment consists of 6 frames, each 64 bytes in size. These segments are transmitted once every 125 microseconds. That means that we have $6 \times 64 \times 8000$ bytes per second or 24.5 Mbit/s of maximum theoretial throughput. But when we looked at the technical specifications for the actual MAN equipment we found that the 34 Mbit/s DSU interface (via an HSSI interface on a cisco running SMDS) has a maximum throughput of 14 Mbit/s (1518 byte packets) while the Ethernet interface had a maximum throughput of 6.46 Mbit/s (also at 1518 byte packets). Our PTT has recommended to us that if we want to maximize the use of the 34 Mbit/s MAN network, we should order multiple Ethernet ports (3) or multiple HSSI ports (2) at each university.

So with 6.46 Mbit/s as the maximum we tried various tests to see if we could attain that speed.

### 4.1. PING

One of the first tests conducted was end-to-end ping tests using various size packets. These tests were repeated hundreds of times during all periods of the day in order to find the "best" time attainable, which are reported in Table 1. The logical ping path was as follows:

local cisco,
MAN switch,
remote cisco,
MAN switch,
local cisco.

We quickly found out that for large packets the round trip time was double that for a LAN. We then had someone in the USA do a test on his 100 Mb/s Fibermux TDM network and he surprisingly got the same numbers as we did. The

Table 2

| Packet size | 1024 | 2048 | 3072 | 4096 | 5120 | 6144 | 7168 | 8096 |
|---|---|---|---|---|---|---|---|---|
| Average packets | 161 | 165 | 155 | 155 | 154 | 133 | 121 | 107 |
| MBits/s | 1.32 | 2.70 | 3.81 | 5.08 | 6.31 | 6.54 | 6.94 | 6.93 |

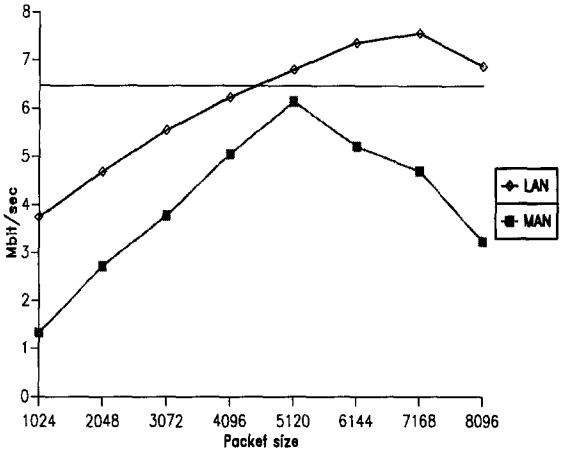


Fig. 3. Comparison of LAN vs. MAN throughput.

explanation cisco found was that the large packets undergo fragmentation/reassembly in the cisco routers and that is what is adding the extra time. So no matter how fast a network one will use, even an ATM-based network running OC-12 circuits, the cisco AGS/4 will prove to be a bottleneck in its handling of large packet sizes.

Next, Oded Comay from Tel Aviv University performed some ping tests using an SGI Challenge. The -f flag was used which floods the network as fast as it can. The buffer size was also played with. Table 2 gives a report of the average number of packets transmitted/received per second.

If one looks carefully at the numbers, you can notice that once packet size becomes larger than 5120 bytes, the packet loss increases, and the number of packets which actually make it all the way over the MAN and back becomes smaller. A better calculation is to count only the received packet rate, and double this. Doing so yields the numbers given in Table 3.

Table 5

| Target workstation | Min. round-trip (ms) |
|---|---|
| Same ethernet | 0.6 |
| Via one cisco | 1.1 |
| Via MAN | 8.7 |

At packet sizes of 5120, we achieve 95% of the maximum throughput Alcatel Bell has stated that an Ethernet MAN can sustain. For a baseline, Oded also did a similar test on a local Ethernet, see Table 4.

These local Ethernet tests show that the peak performance for an Ethernet (our Ethernet in this case) is around 7.57 Mbit/s when packet size is 7168 octets, see Fig. 3. A quick look suggests that for reasonable size packets (1500 octets), MAN is 2–3 times slower than a local Ethernet. For large packets—those around 5120 octets (being fragmented at a level below the application)—MAN performance is comparable to local Ethernet. This suggests that a problem exists with the MAN, not of throughput, but latency, which mostly affects small packets.

### 4.2. Latencies

The next set of tests was to determine the latencies induced by the MAN network. Oded Comay performed another set of tests but this time with a Network General Sniffer, who's accuracy is $\pm 0.1$ ms. Table 5 shows the minimal round-trip time taken for a minimal frame to go from one workstation to another.

Table 3

| Packet size | 1024 | 2048 | 3072 | 4096 | 5120 | 6144 | 7168 | 8096 |
|---|---|---|---|---|---|---|---|---|
| Received rate | 81 | 83 | 77 | 77 | 75 | 53 | 41 | 25 |
| Mbits/s | 1.33 | 2.72 | 3.78 | 5.05 | 6.14 | 5.21 | 4.70 | 3.24 |

Table 4

| Packet size | 1024 | 2048 | 3072 | 4096 | 5120 | 6144 | 7168 | 8096 |
|---|---|---|---|---|---|---|---|---|
| Received rate | 229 | 143 | 113 | 95 | 83 | 75 | 66 | 53 |
| Mbits/s | 3.75 | 4.69 | 5.55 | 6.23 | 6.80 | 7.37 | 7.57 | 6.87 |

Table 6

| Target | Min. round-trip (ms) |
|---|---|
| Tel Aviv MAN | 1.2 ← local |
| Bar-Ilan MAN | 8.2 |
| Weizmann MAN | 8.4 |
| Hebrew U. MAN | 9.6 |

Next was measured the minimal round-trip time from the workstation to the various MAN interfaces. The results are given in Table 6.

From these ping timings one can deduce that minimal round-trip over the MAN (removing Tel Aviv University internal latencies) is as low as 7 ms (to Bar Ilan University—located on the same MAN switch as Tel Aviv University) up to 9.6 ms (to Hebrew University—located on a MAN switch 60 km away). The conclusion on latency timings within the MAN switches would then be as shown in Table 7.

Table 7

| | |
|---|---|
| 3.5 ms | if traffic stays within a single switch |
| 4.2 ms | if traffic traverse a single hop |

In [1], Steven Taylor says in regards to windowing:

But a more serious performance problem occurs when sizes are set too low. If the last frame in a window is sent before the first frame has been acknowledged, the sending device must wait until that acknowledgement is received before transmitting more data. This "windowing out" can create chronic performance problems on high speed links. In fact, the faster the WAN circuit, the more acute the windowing-out problem becomes...

One way to minimize windowing out is to neutralize the transmission delay factor by making sure that the total number of bytes in a window (the number of frames multiplied by the number of bytes in each frame) is greater than the amount of traffic in the transmission pipe at any given time.

In our case, the transmission pipe is rated at 6.46 Mbit/s and the round trip latency is approximately 9 ms. This translates into 60 964 bits or 7620 bytes that can be inside the pipe at any one time.

If one uses a window of 8 and a packet size of 1024, then the pipe will be used effectively. But IPX, for example suffers from a small packet size and a very poor WAN protocol. A detailed test of IPX was done to verify this.

### 4.3. IPX

IPX is a protocol extremely susceptible to WAN latencies. This is because Netware 3.X used a packet size of 576 octets and a window size of 1. Later versions of Netware (such as 4.0) used a packet size of 1500 but unless PBURST is used (which used a window of 16) the window is retained at one. This set of tests were performed by Doron Shikmoni and Gershon Kunin of Bar-Ilan University.

A 486 PC was set up behind the cisco router at Bar-Ilan University. The other side was a Netware server at Weizmann Insitute of Science which ran a Netware server (located behind 2 cisco routers). RAM disks were used as well as

Table 8

| | | | |
|---|---|---|---|
| A. | Workstation with NETX shell (no Packet Burst) | | |
| | MAN: | Read: 487.2 Kb/s | Write: 485.6 Kb/s |
| | Local: | Read: 1360.0 Kb/s | Write: 1398.4 Kb/s |
| B. | Workstation with VLM, PB BUFFERS = 0 (no Packet Burst) | | |
| | MAN: | Read: 464.8 kb/s | Write: 460.8 kb/s |
| | Local: | Read: 1176.0 kb/s | Write: 1186.4 kb/s |
| C. | Workstation with VLM, PB BUFFERS = 3 (default: Packet Burst active) | | |
| | MAN: | Read: 2815.2 kb/s | Write: 1893.6 kb/s |
| | Local: | Read: 4556.8 kb/s | Write: 3376.0 kb/s |

Netware cashing. This was not a lab test to see how fast IPX can run but rather how fast this particular test setup could run. At Bar-Ilan Netware was 3.12 and at Weizmann it was 4.01. The results are given in Table 8.

### 4.4. FTP

The trick to performing proper FTP benchmarks is to eliminate as much as possible from the surrounding environment. That includes minor things like disk seeks, Ethernet contention and memory allocation. The person who performed some of the FTP benchmarks was Yaron Zabary at Tel Aviv University. A 4 Mbyte file was FTPed to a remote MAN site. The file was sent to /dev/null so that at the remote site there would be no disk seeks. In order to eliminate disk seeks locally, the test was repeated a number of times, since on the first time through, the Unix system reads the file to memory but on all subsequent executions, the file resides in cache.

The best time achieved was 517.61 Kbytes/s which came to just over 4 Mbit/s. We used an SGI to SGI system for the FTP but did not use large window (RFC1323) support. No attempt was made to see how throughput would be affected by many parallel FTP sessions.

## 5. Conclusions

The MAN network based on Ethernet ports is extremely simple to install. Its throughput is not quite 10 Mbit/s but it is far better than an E1 (2 Mbit/s) digital circuit. An Ethernet "passthrough" port does provide an evolutionary path for those who wish to migrate from fractional E1 lines to ATM speeds (E3: 34 Mbit/s) but can't yet afford the high costs of such circuits. Sites are recommended to exercise caution when accepting new services of this nature and to demand benchmarks of throughput and latency from their telecommunications service providers.

## Acknowledgements

I would like to thank all the Israeli university networking personnel that worked hard to test and debug this network: Oded Comay, Doron Shikmoni, Simon Shikman, Benzi Mizrachi and Rafi Sadowsky.

## References

[1] S. Taylor, Moving to faster pipes? Watch those protocols, *Data Comm. Mag.* (September 1994).

**Hank Nussbacher** is a networking consultant who works for the Israeli InterUniversity Computer Center and often acts as a crach dummy on the Israeli Information Superhighway.