LAN and WAN Tests with Objectivity 5.1

Monarc Test-beds Working Group

A.Brunengo (INFN Genova)

A.Ghiselli (INFN Cnaf)

L.Luminari (INFN Roma1)

L.Perini (Milano University and INFN Milano)

S.Resconi (CILEA and INFN Milano)

M.Sgaravatto (INFN Padova)

C.Vistoli (INFN Cnaf)

1 Introduction

This note documents some preliminary network tests with Objectivity 5.1 for MONARC test-beds. The work is still in progress and further developments are foreseen.

The tests have been designed having in mind the request to measure network parameters for the model simulation starting from simple scenarios. Therefore different network scenarios have been set up consisting of a single federated database, one AMS server and several clients locally or geographically distributed. Only read operations are allowed to the clients.

Measures of CPU utilization on server/client workstations and network throughput with different number of jobs have been collected and discussed. Future test scenarios have been proposed.

2 Test Description

Network environment can affect client/server systems for several reasons:

1. Overhead due to communication protocols. For example network throughput can change significantly modifying TCP flow control parameters.
2. Application protocols (how client/server exchange data)
3. Network speed and system capability to use it.
4. End-to-end delay and relationship with link speed and throughput.

The tests described in this document are significant concerning point 3 and 4. In order to investigate points 1 and 2 it is necessary to approach Objectivity architecture and software implementation.

Tests are based on several client/server configuration over different LAN and WAN scenarios with network speed ranging from 2Mbps up to 1000Mbps. Test results have been compared and discussed.

2.1 Objectives

The most important specific objectives are:

- Check AMS behavior and performance.
- Stress tests by running several analysis jobs accessing to the Data Base and performance measure.
- Locate system bottlenecks.
- Collect 'response time' measures to give input to model simulation.
- Understand network traffic characteristics and profiles.

2.2 Test Characteristics

The general test scenario is very simple regarding to database characteristics and structure. This choice has been done in order to evaluate the system performance in the simplest configuration.

- ATLFast++ program [1] is used to populate Objectivity database following the Tag/Event data model proposed by the LHC++ project. The database architecture is very simple: one single container for the event and no associations in the database.

- 1 single Objectivity federation containing about 50000 events, corresponding to about 2 GB (the event size is about 40 KB).
- Application program performs only read access to the database. The jobs read about 3000 events each (about 120 MB) except the Babar farm case where the jobs read 10000 events (about 400 MB).
- Stress tests have been performed. The procedure followed consists in submitting an increasing number of concurrent jobs from each client and then monitoring CPU utilization, network throughput and job execution time (wall clock time). The same kind of tests have been performed on a local federated database [2].
- System configurations considered are: one server / one client and one server / many clients.
- Objectivity 5.1 has been used setting the page size at 8192 bytes.
- 1 AMS server per scenario.

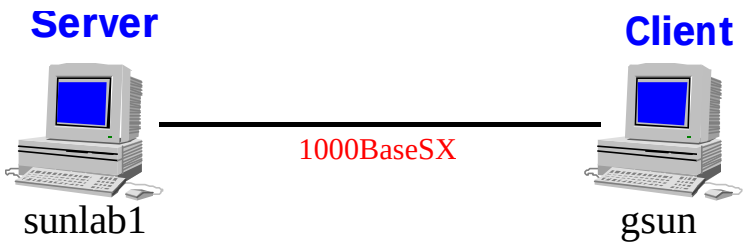
The network setups are with various bandwidth starting from 2Mbit/sec (WAN) up to 1000 Mbit/sec (LAN):

- LAN at 10Mbps, 100Mbps and 1000Mbps
- WAN scenario in production environment.

Details of the performed tests are described in the following; they are grouped in 'one server/one client' and 'one server/several clients'.

3 One Server / One Client Tests

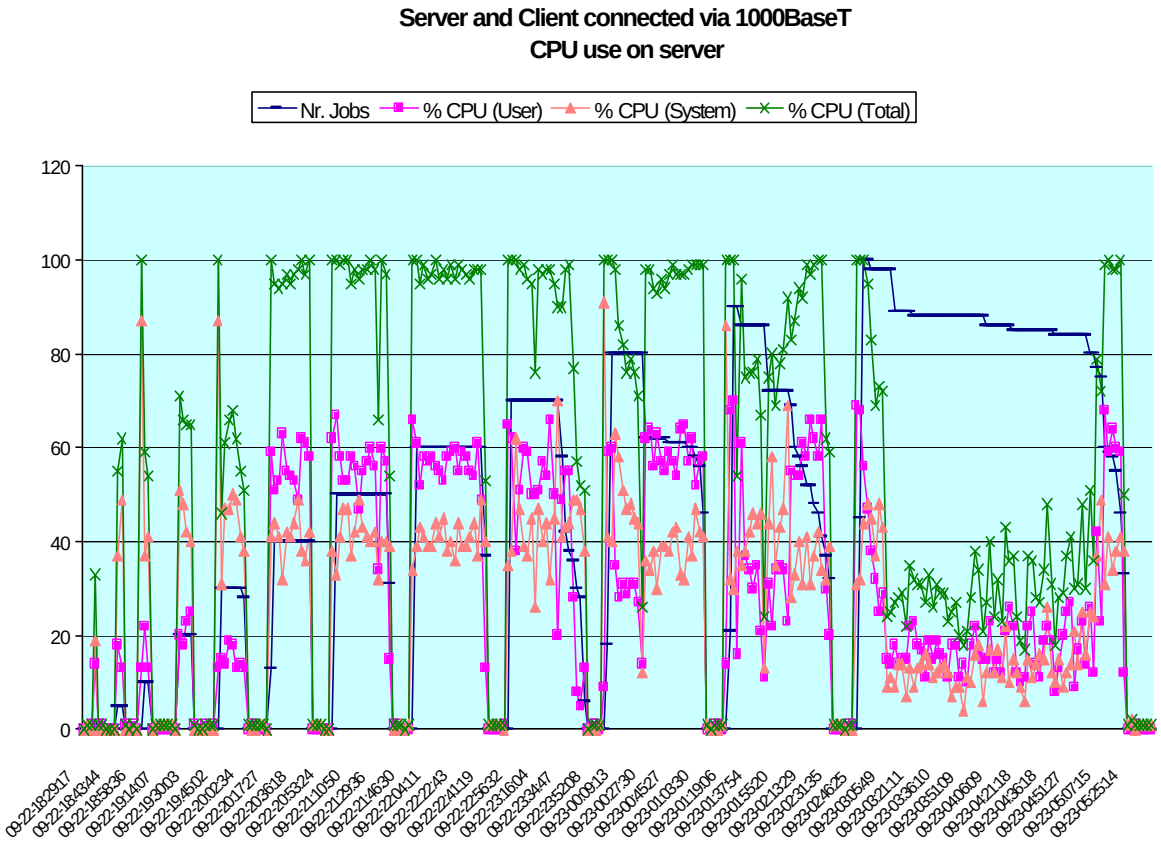
3.1 LAN Test 1

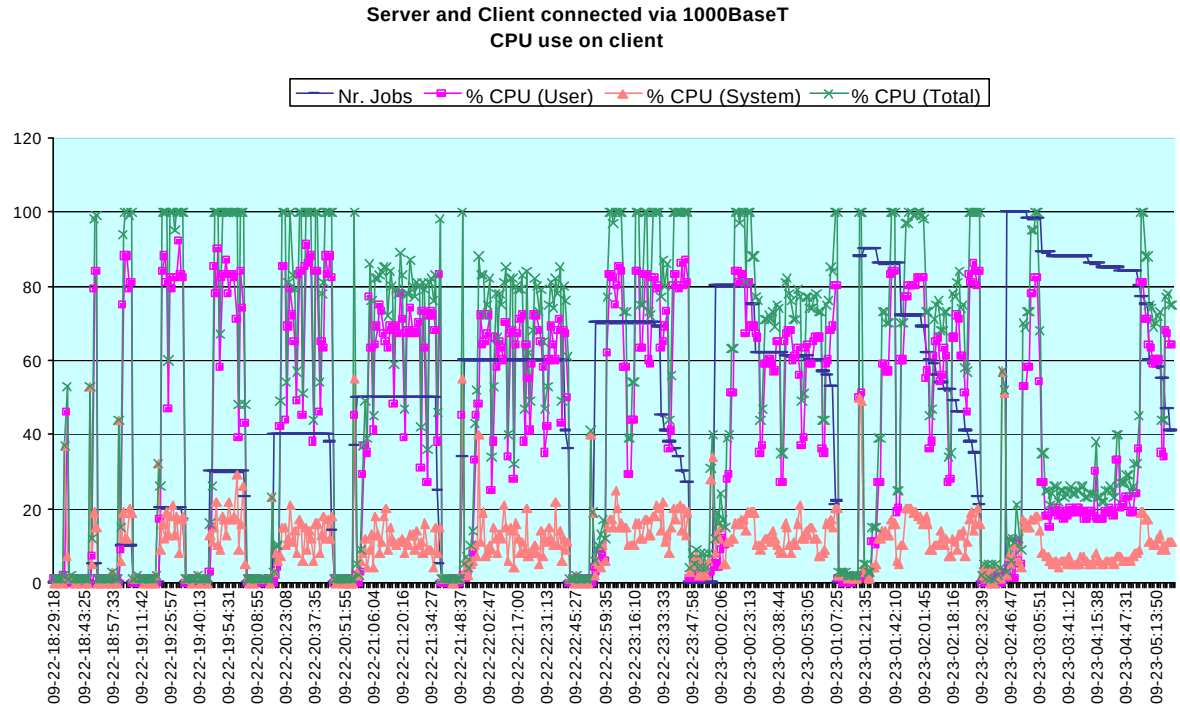


3.1.1 Hardware Configuration

	CPU	Memory	Disk	OS	Link
Sunlab1	ULTRA5, 333MHz	128MB	SCSI3 2*9GB	Solaris 2.7	Sun PCI GE Multimode fiber
Gsun	ULTRA5, 333MHz	128MB		Solaris 2.7	Sun PCI GE Multimode fiber

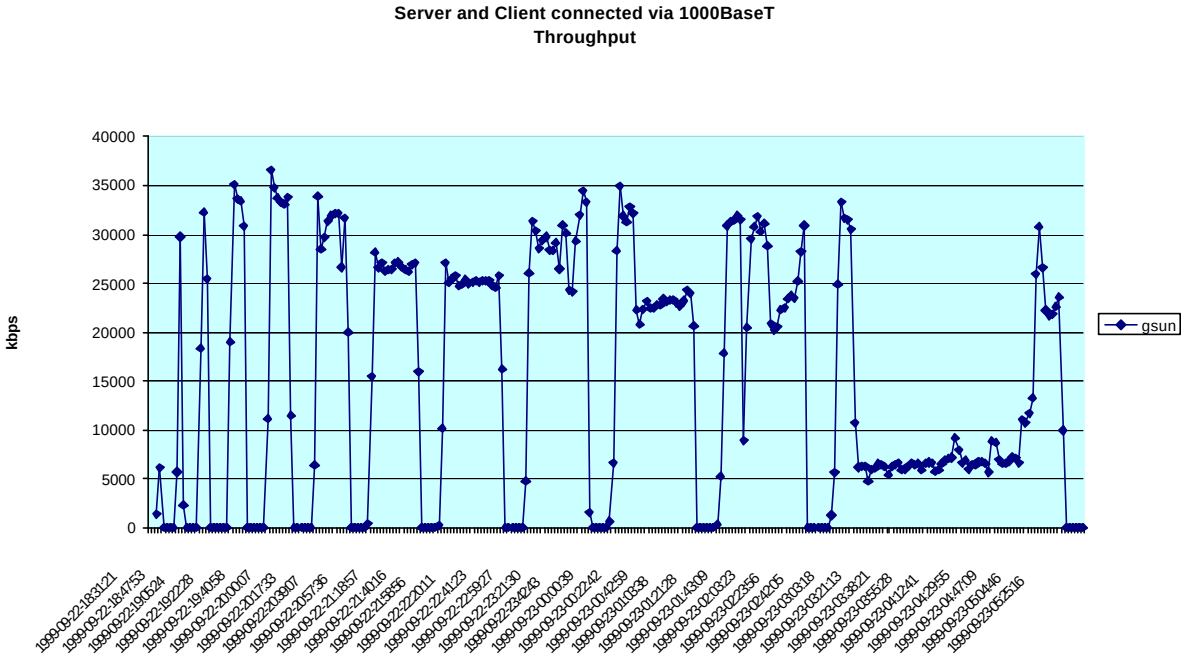
Plots of the CPU use on server and client machine follow. The two figures below show that the CPU usage on client machine is highest when few jobs are running, and it stays at this level until the server reaches 100% of CPU usage. Then server CPU use remains at 100% whereas on client the CPU use decreases. In the same timeframe, throughput behavior between the two machines shows a maximum with 40 concurrent jobs, then it decreases. In particular it is interesting to point out that the client CPU utilization reaches 100% after 5 concurrent jobs and the server CPU utilization reaches 100% after 50 concurrent jobs. After 80 concurrent jobs on client, some jobs start to crash.





3.1.2 Throughput

Network throughput has been measured on the server counting user-data of each network send operation.



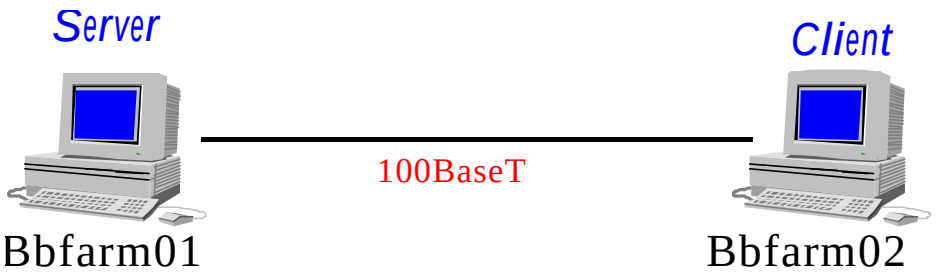
The figure shows that the maximum throughput value is 37 Mbps corresponding to 30 concurrent jobs and when corresponding CPU use on server is still below 100%.

3.1.3 Job Wall Clock Time

Job wall clock time (WCT) on client machine is estimated as average value for each set of concurrent jobs.

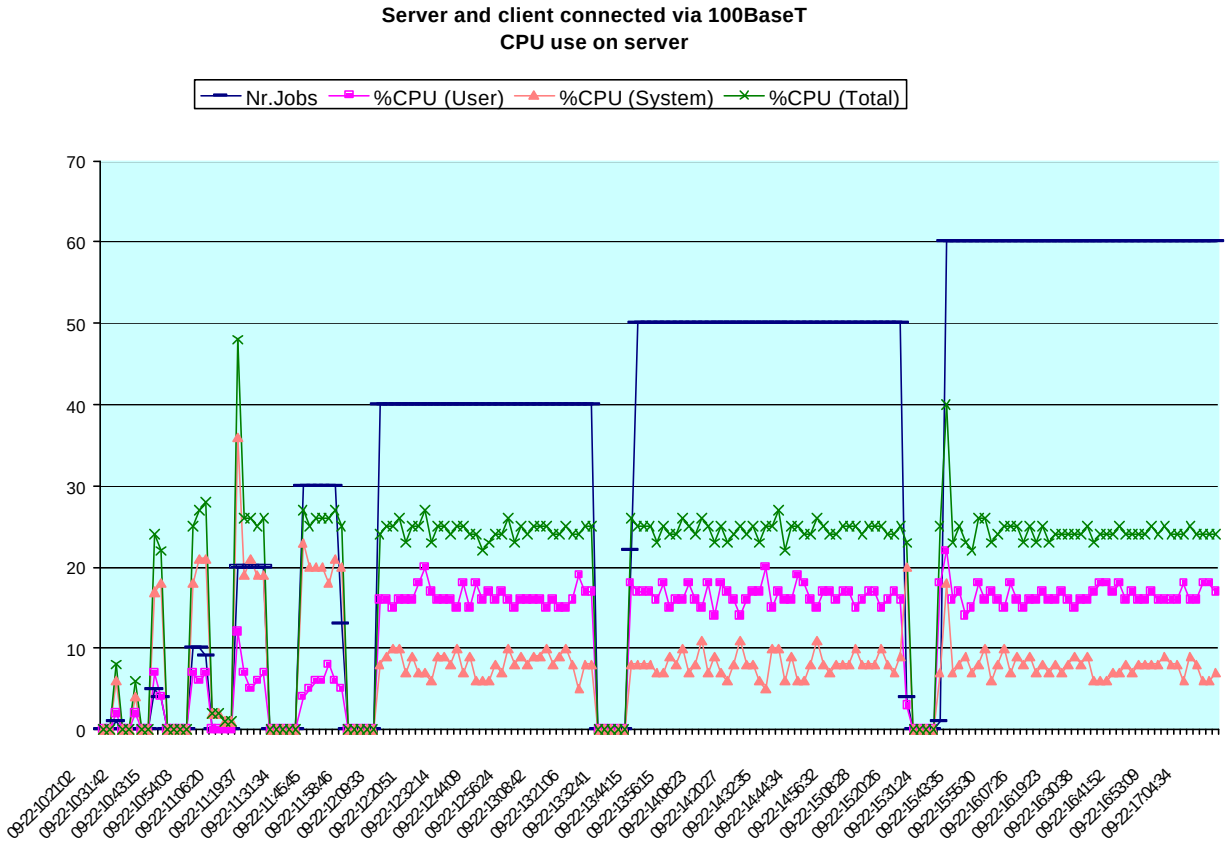
WCT changes significantly after 30 jobs and keeps the same value up to 60, then the situation is very unstable. After 40 concurrent jobs, server CPU is saturated, throughput decreases and WCT increases but not in an homogeneous way.
As general comment we can say that the CPU power of Sun Ultra5 is the bottleneck of the system (client and server CPU saturate very quickly) and it is completely inadequate for a 1000Mbps (only 30Mbps of throughput).

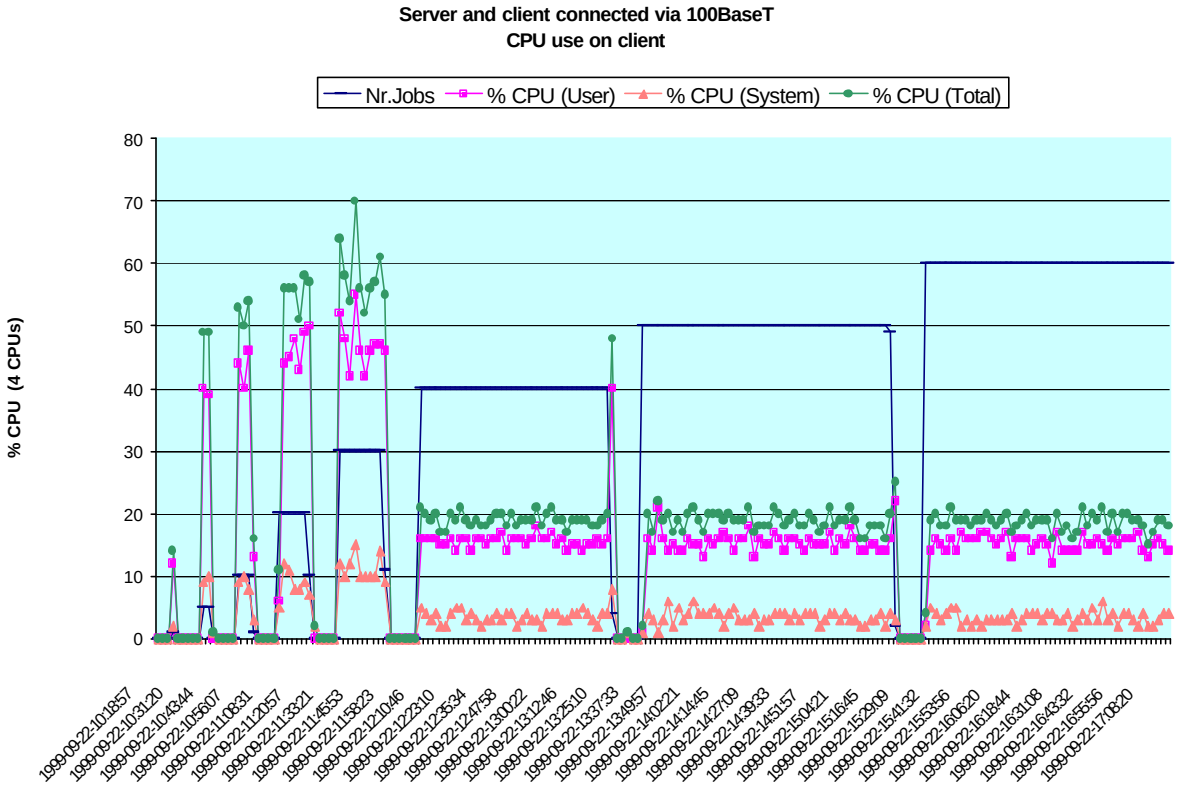
3.2 LAN Test 2



	CPU	Memory	Disk	OS	Link
Bbfarm01	Sun E450 Ultra 2 400MHz, 4 CPU	512MB	RAID A3500	Solaris 2.7	Sun PCI FE
Bbfarm02	Sun E450 Ultra 2 400MHz, 4 CPU	512MB		Solaris 2.7	Sun PCI FE

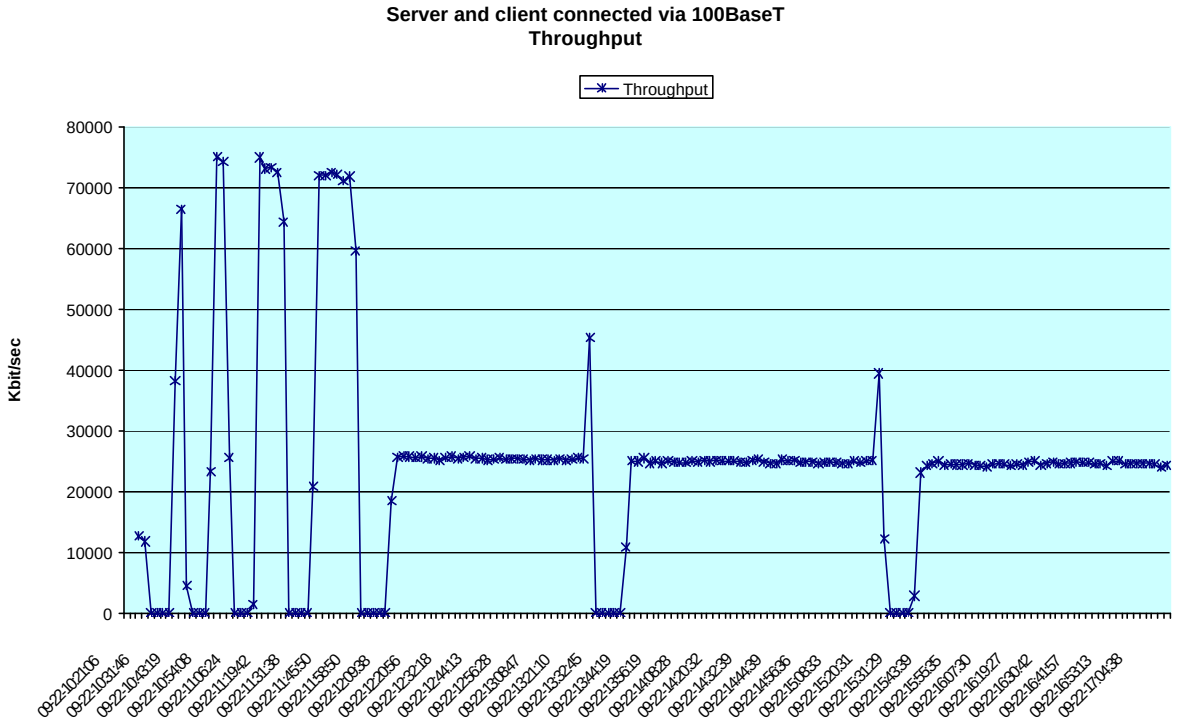
3.2.1 CPU Use


Server and client connected via 100BaseT
CPU use on server
Nr.Jobs
%CPU (User)
%CPU (System)
%CPU (Total)
70
60
50
40
30
20
10
0
09.22.102102
09.22.103142
09.22.104315
09.22.105403
09.22.110620
09.22.111937
09.22.113134
09.22.114545
09.22.115846
09.22.120933
09.22.122051
09.22.123214
09.22.124409
09.22.125624
09.22.130842
09.22.132106
09.22.133241
09.22.134415
09.22.135515
09.22.140823
09.22.142027
09.22.143235
09.22.144434
09.22.145632
09.22.150828
09.22.152026
09.22.153124
09.22.154335
09.22.155530
09.22.160726
09.22.161923
09.22.163038
09.22.164152
09.22.165309
09.22.170434



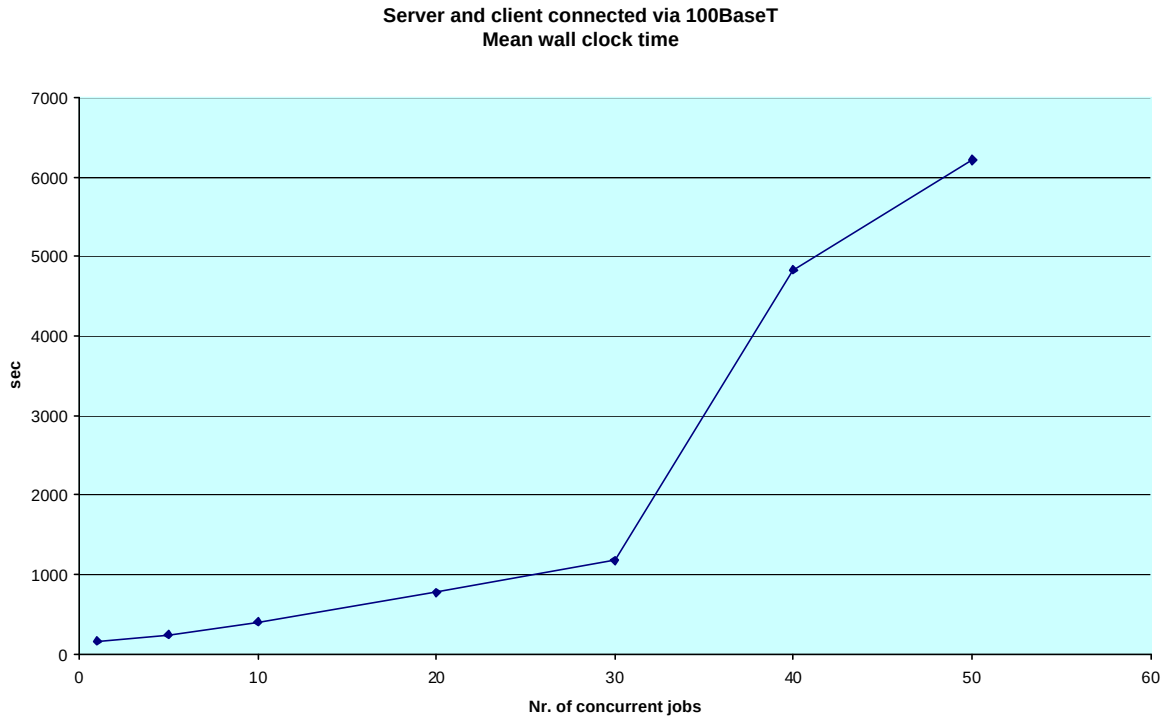
Client CPU use for 5 jobs is very high (50 %) of a 4*CPU machine. It is interesting to point out that up to 30 concurrent jobs CPU usage on client is around 60 % and for an higher number of jobs it decreases to 20. In the same range of concurrent jobs server CPU use is 100 % and network throughput is nearly 100Mbps, as shown below.

3.2.2 Throughput



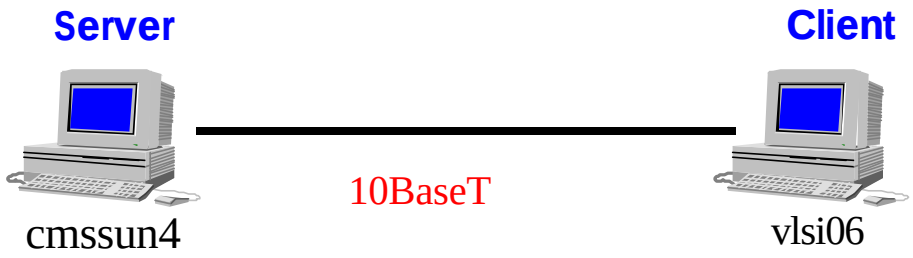
The sharp throughput drop corresponds to CPU saturation on server and to client CPU decreasing. The 'sharp' drop has to be investigated.

3.2.3 Job Wall Clock Time



This figure confirms the discontinuity at 30 jobs. As general comment we can say that the AMS 5.1 limitation on using multiprocessor systems is the bottleneck of the system.

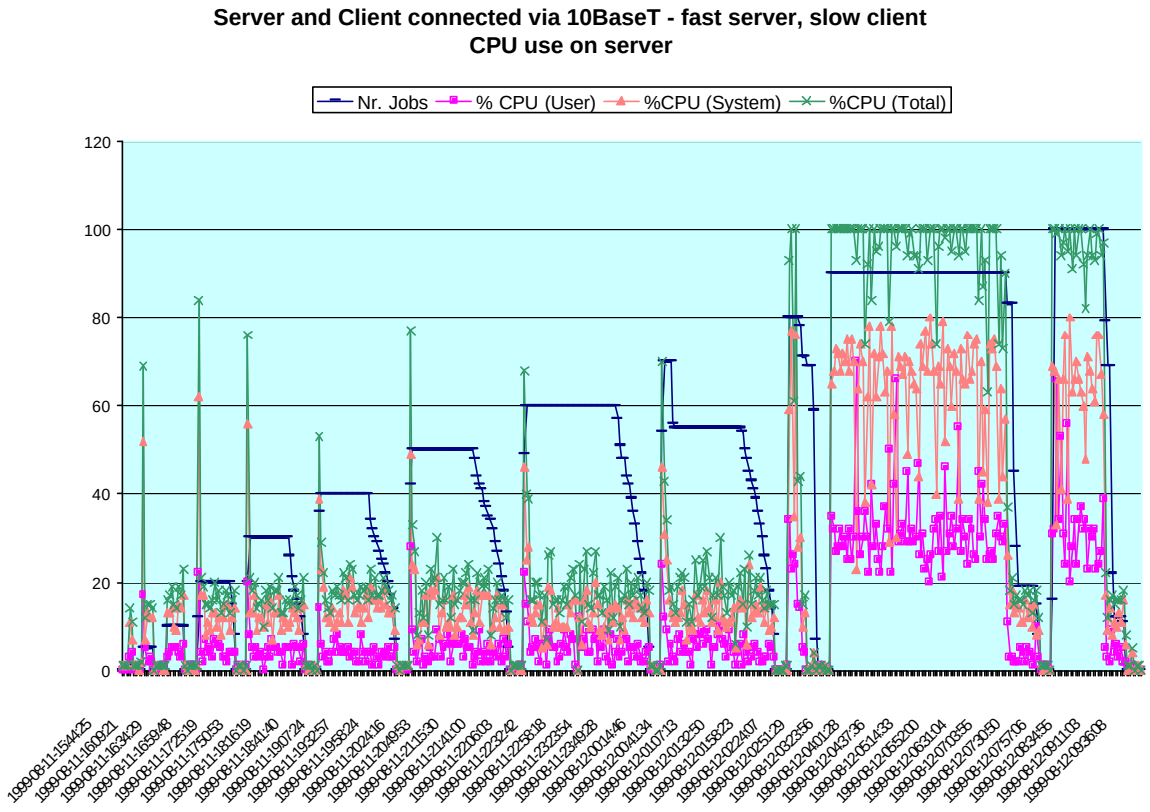
3.3 LAN Test 3

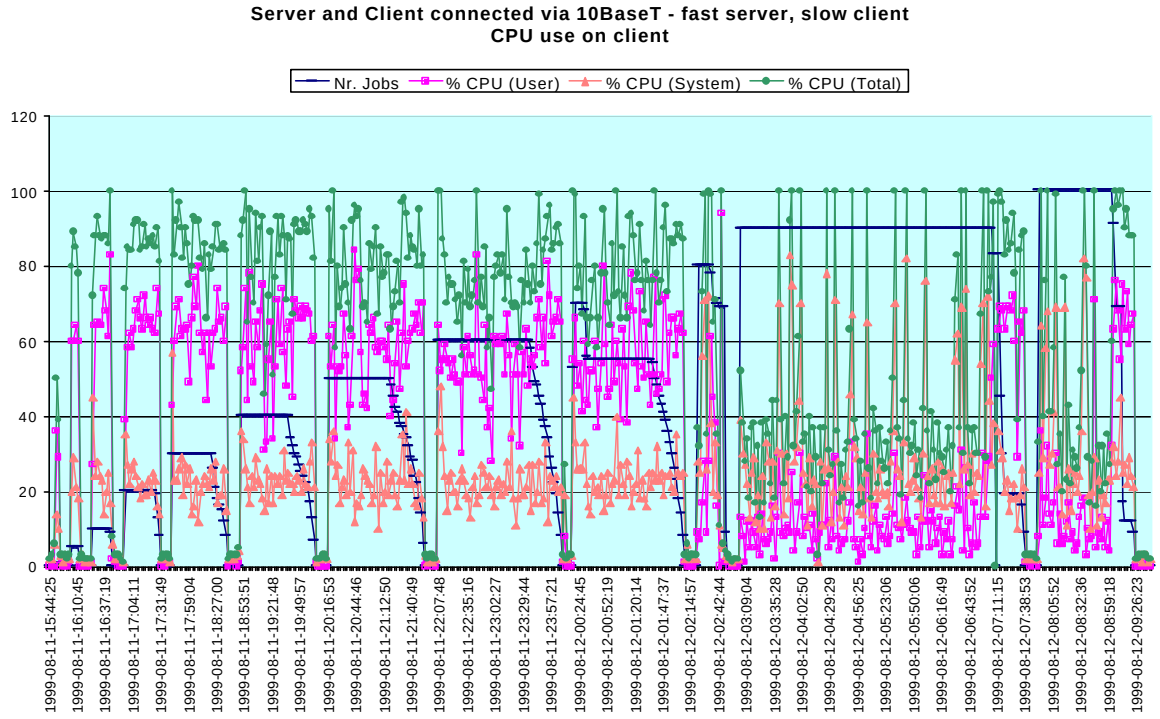


3.3.1 Hardware Configuration

	CPU	Memory	Disk	OS	Link
Cmssun4	SUN ULTRA10, 333MHz	128MB	SCSI3 2*9GB	Solaris 2.6	10M Ethernet interface
Vlsi06	Sun SPARC20, 125Mhz	128MB		Solaris 2.6	10M Ethernet interface

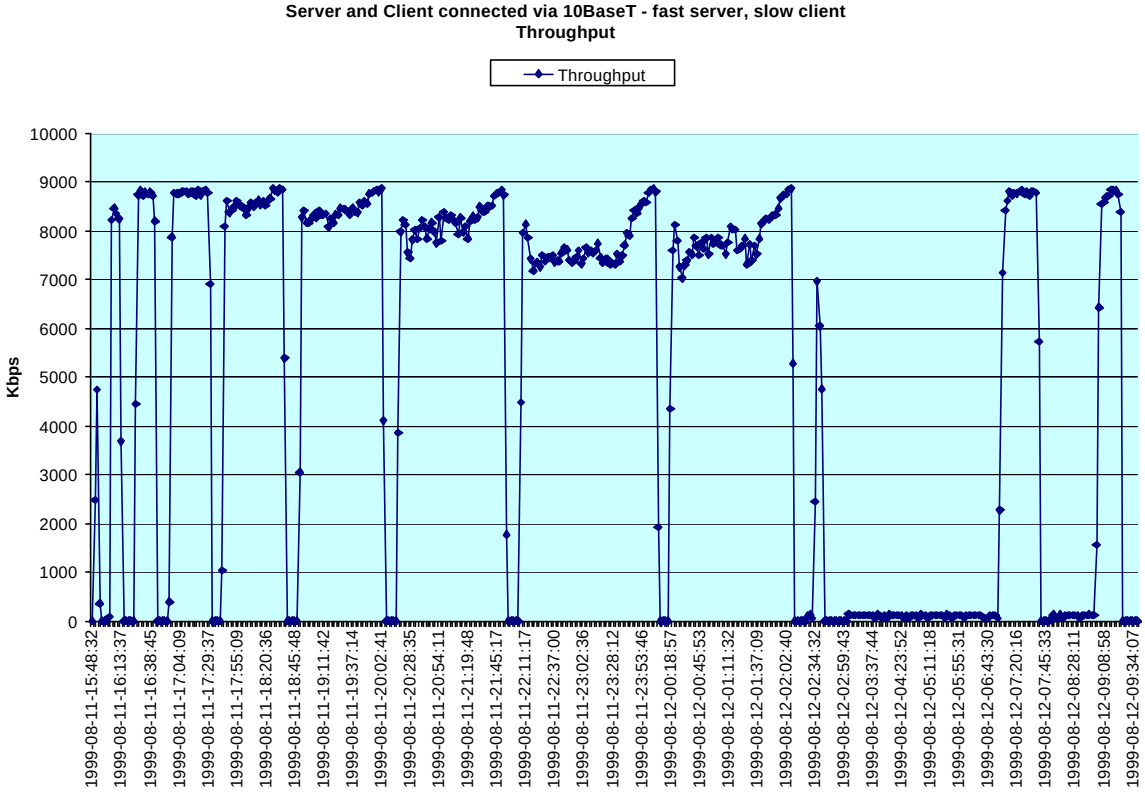
3.3.2 CPU Use



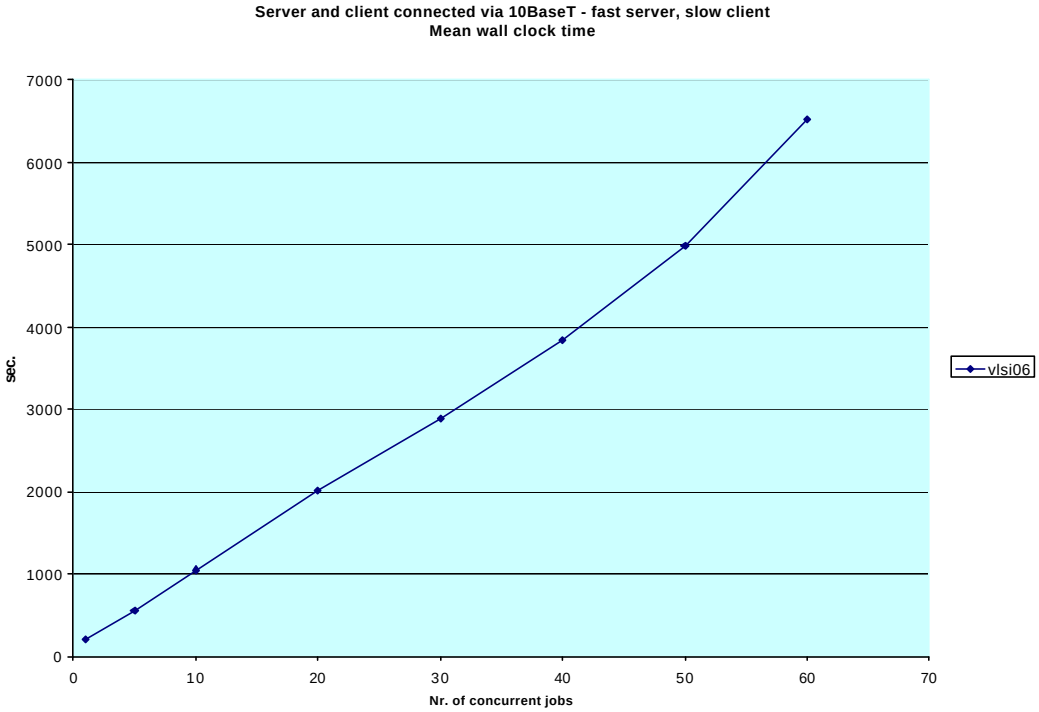


It is interesting to notice that client CPU utilization after 20 jobs and up to 70 is in the range 80-100 %. With 70 jobs some of them start to crash. The server CPU utilization is less than 30 % up to 60 jobs and when jobs crash on the client, CPU utilization reaches 100 %.

3.3.3 Throughput

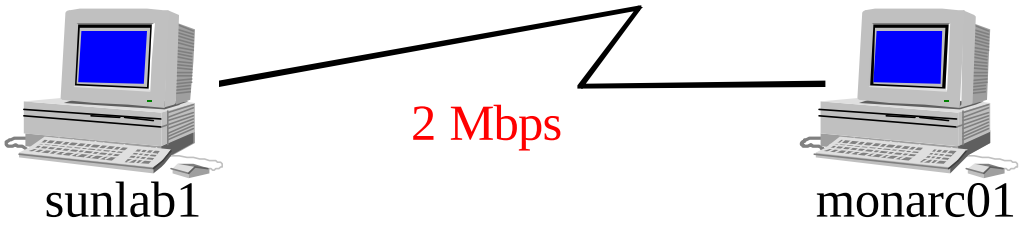


3.3.4 Job Wall Clock Time

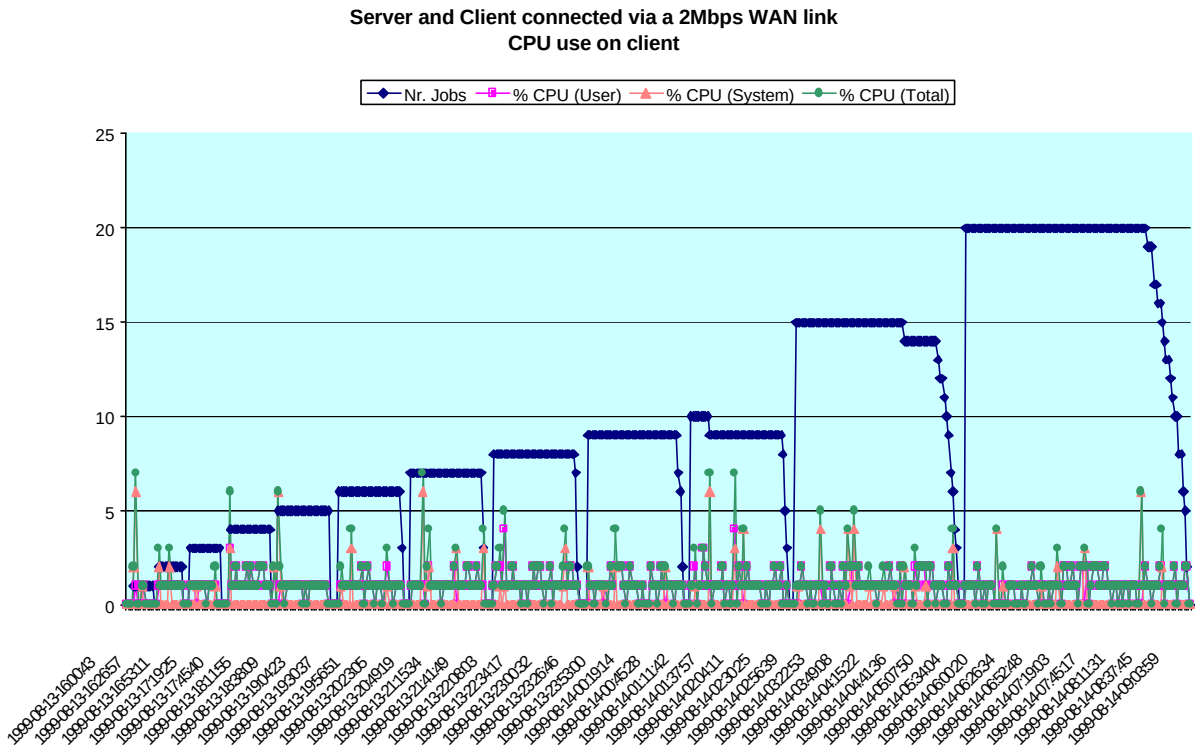


10 Mbps speed link is well utilized and it does not saturate server CPU. The less powerful client CPU saturates very quickly causing a throughput decrease. It is interesting to notice that when job start crashing something happens on the server that increases CPU up to 100 %.

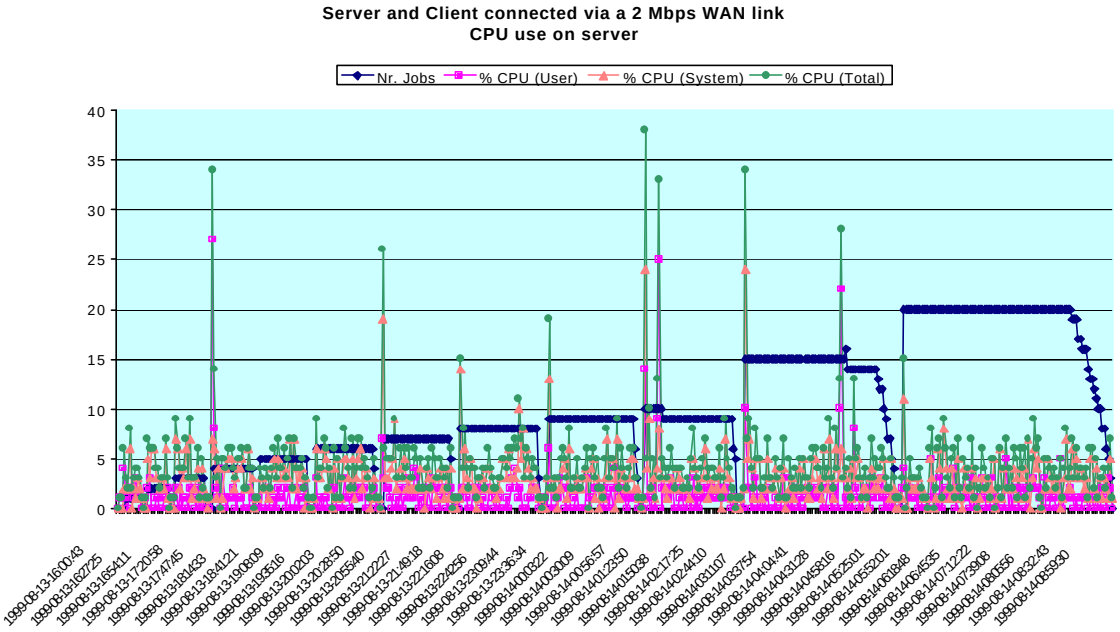
3.4 WAN Scenario Test 4



	CPU	Memory	Disk	OS	WAN Link: ATM CBR VC at 1.7Mbps
Sunlab1	Sun ULTRA5, 333Mhz	128MB	SCSI3 2*9GB	Solaris 2.7	
Monarc01	Sun Enterprise 450, 4*400Mhz	512MB		Solaris 2.6	

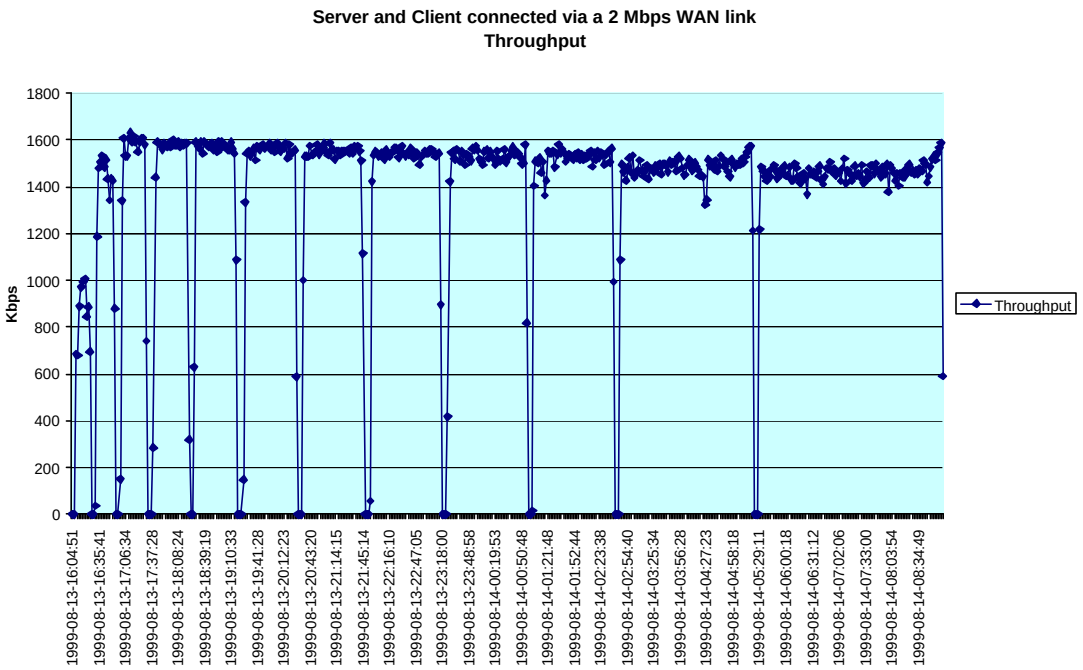


This test has been done over a dedicated 2Mbps link. Since the link is based on ATM VP, the real payload is 1.7Mbps (4717cells/sec→ 4717*48bytes/sec = 1811Kbps – AAL5&snap overhead [16bytes per PDU] = 1800Kbps – IP overhead [40bytes per PDU] = 1700 Kbps).

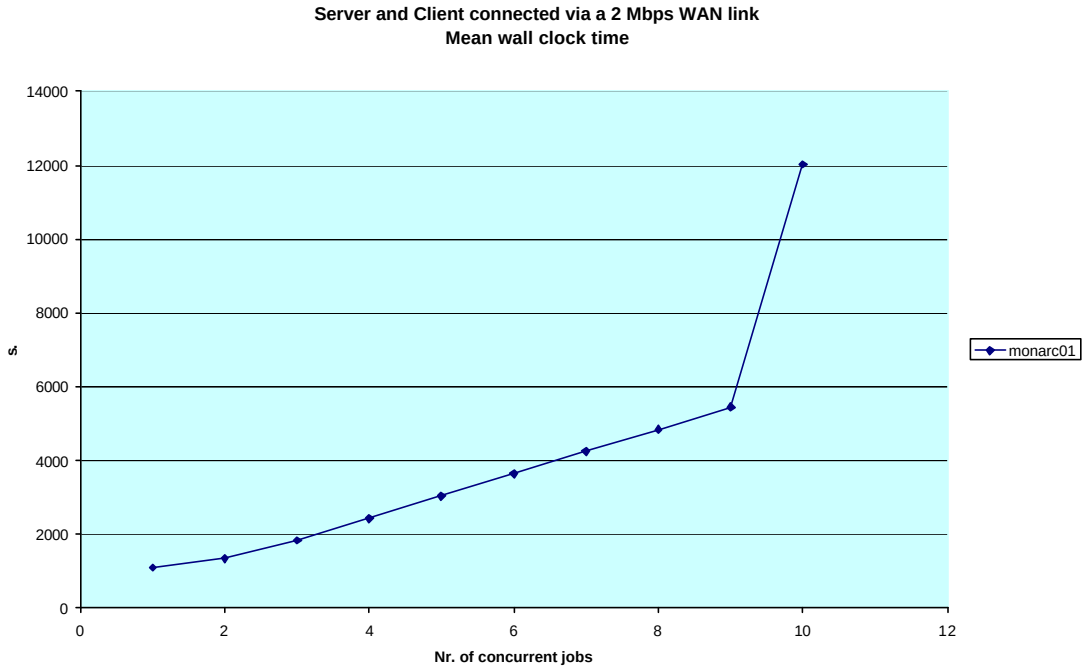


Client CPU utilization is always less than 5 % and the server CPU utilization is less than 10 %. It has to be investigated the reason why with 10 and 15 concurrent jobs one crashes.

3.4.1 Throughput



3.4.2 Job Wall Clock Time



Mean wall clock time is very high: with 5 concurrent jobs on a client, mean WCT is 3000 sec whereas the correspondent value in test 1 is 500 sec.

3.5 One Server / One Client Test Summary

3.5.1 CPU

The following table summarizes server and client CPU utilization versus network speed, with the corresponding running jobs.

	CLIENT		SERVER	
Network speed	Max CPU	Number of jobs running	Max CPU	Number of jobs running
1000M	100%	≥5	100%	≥50
100M	60% , then 20%	Up to 30, then up to 60	100%	Up to 60
10M	80%	≥20	30%	≥60
2M	5%	Up to 20	10% (constant)	1-20 (during the all test)

The above table shows that:

- 1000 Mbps, CPU client is the bottleneck
- 100 Mbps, CPU server is the bottleneck
- 10 Mps and 2 Mbps, network is the bottleneck

3.5.2 Throughput

The following table summarizes the maximum throughput obtained versus line speed and running jobs:

Link	Server host	
Network speed	Max throughput	Number of jobs
1000M GEthernet	37Mbps	≥20
100M FEthernet	80-100Mbps	≤30
10M Ethernet	9Mbps	≥20
2M VC ATM	1.7Mbps	≥20

Network utilization is good for 10BaseT and for 2M ATM VC; in case of Gigabit Ethernet network utilization is very low and it must be investigated with future release of Objectivity and with more powerful machines.

3.5.3 Job wall clock time

In order to compare the job execution times between the tests, an average wall clock time for one job has been calculated taking into account the 10 job measures:

- Test 1, 1000 Mbps, mean wall clock time 360 sec, single job 60 sec.
- Test 2, 100 Mbps, mean wall clock time 150 sec, single job 48 sec.

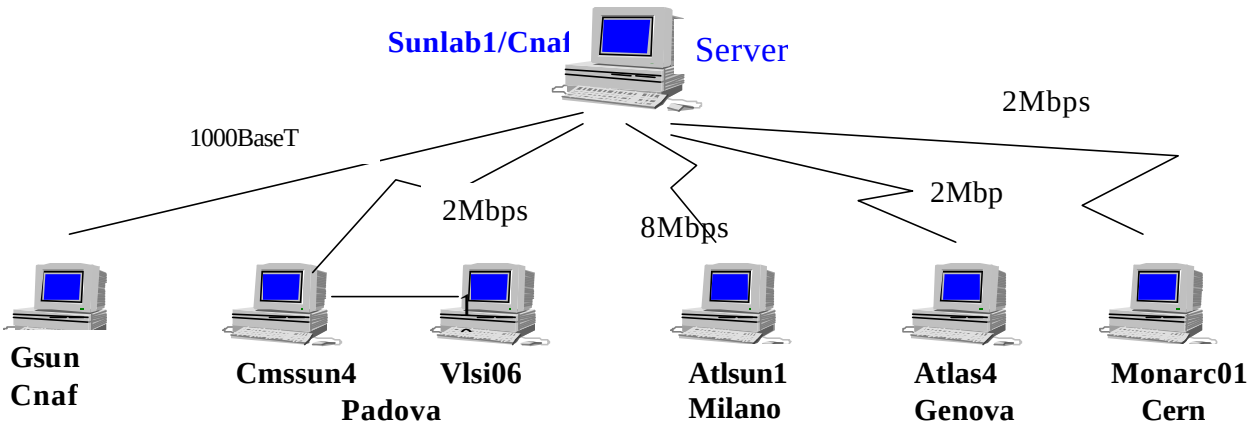
- Test 3, 10 Mbps, mean wall clock time 1000 sec, single job 200 sec.
- Test 4, 2 Mbps, , mean wall clock time 6000 sec, single job 1000 sec.

It could be interesting to enhance that, with the same CPU power conditions, wall clock times, from Test 1 up to Test 4, increase with the same factor as throughput (as it was expected): Test 3 wall clock time is 2.5 times Test 1 wall clock time and that there is the same factor between Test 1 throughput and Test 3 throughput. Test 4 wall clock time is 6 times Test 3 wall clock time and Test 3 throughput is 5.6 times Test 4 throughput. Test 2 is an exception since it is based on different Sun architecture.

4 One Server / Several Clients Tests

4.1 Wide Area Network Test

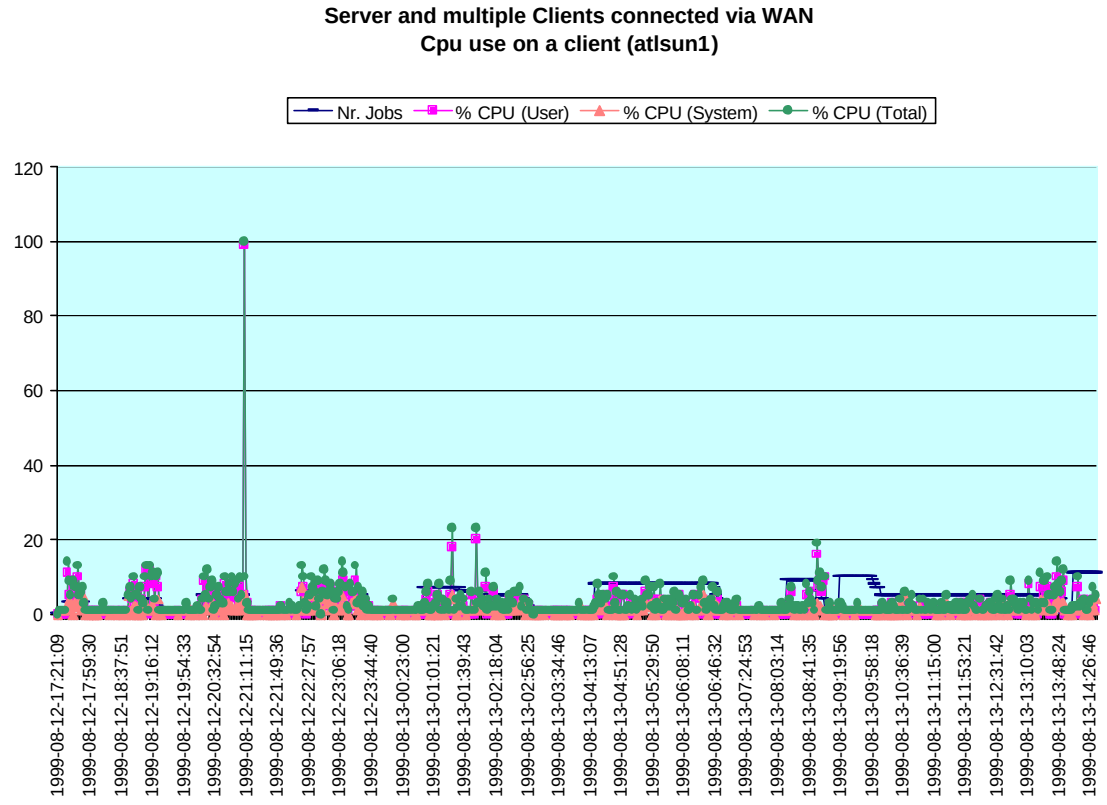
A script on each client runs an increasing number of jobs starting from one.



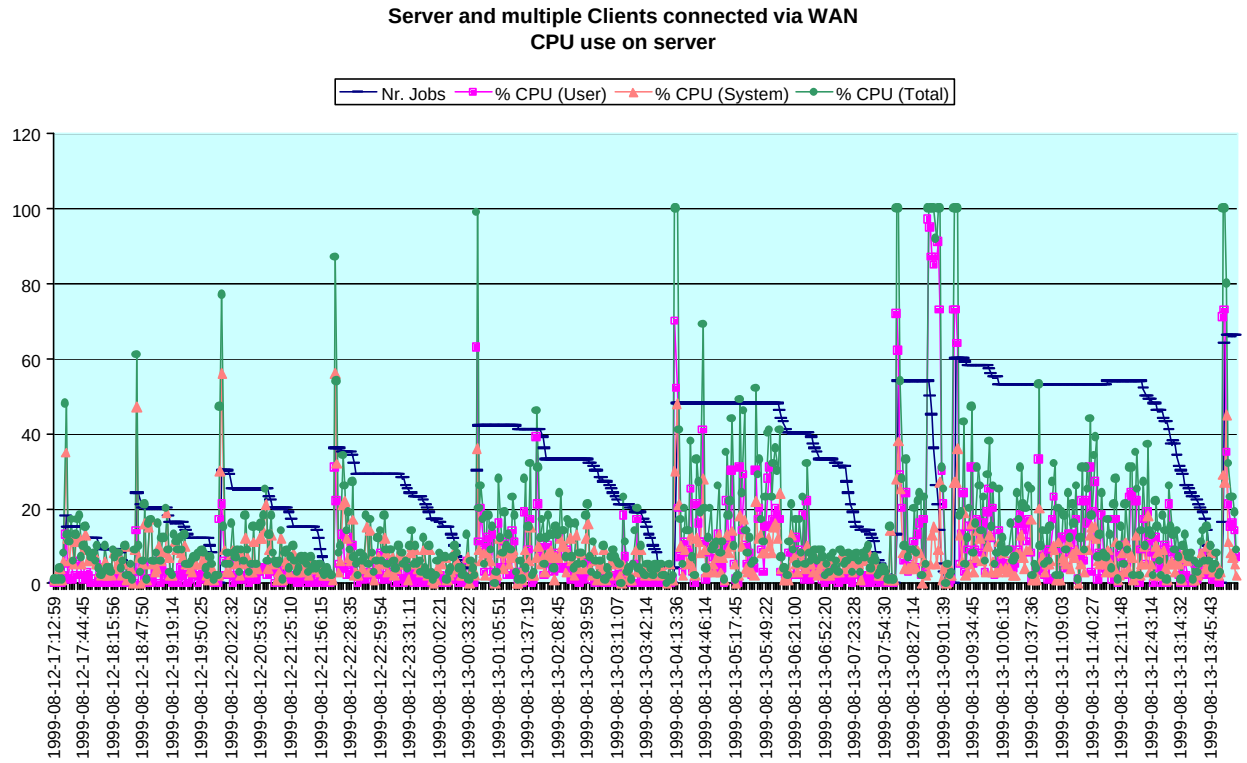
4.1.1 Link Description

Milano, Genova, Padova are connected to CNAF through production network and CERN is connected to CNAF through a dedicated ATM VC.

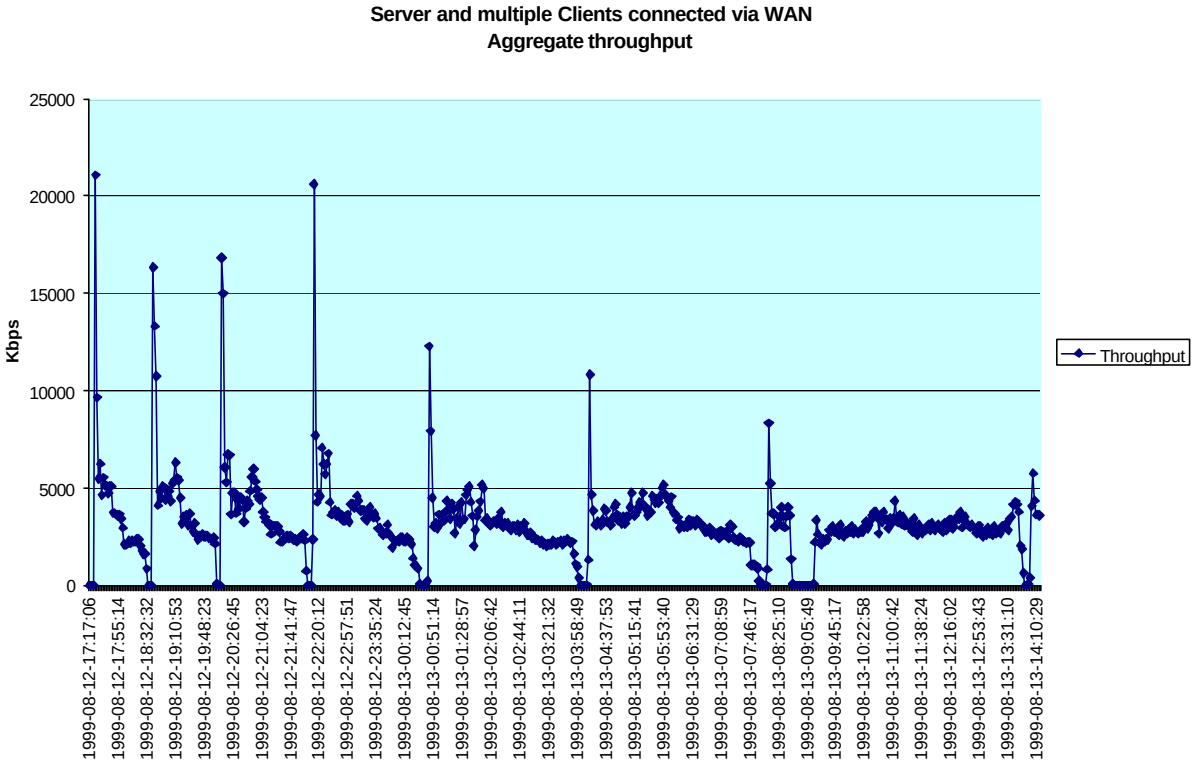
	Access Link Technology to GARR-B (ATM backbone)	Access Speed to GARR (Mbps)	RTT to Cnaf
Cnaf	ATM	6	
Padova	Point-to-point	2	7 msec
Genova	Frame-relay	2	21 msec
Milano	ATM	8	8 msec
Cnaf - Cern	ATM	2 (4717 cells/sec)	51 msec

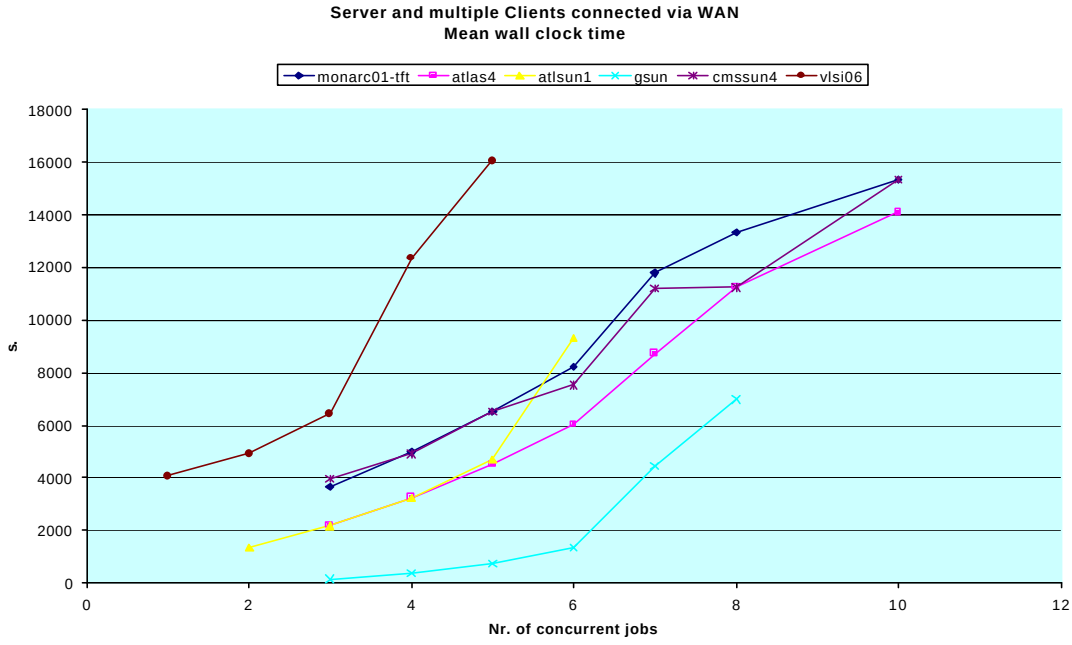


The above figure shows the behavior of CPU utilization on one of the 6 clients as example. In all of them CPU usage is always under 20 %.



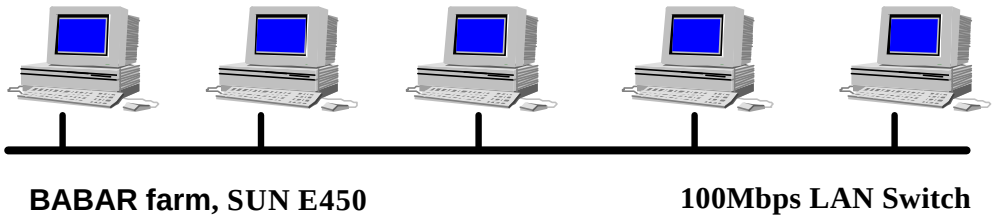
CPU utilization on the WAN server is rather small, under 40 % with 50 jobs. Using 25 sec as RPC timeout some jobs crash after 30 on server: for example with 6 jobs per client, 2 jobs crash on vlsi06; with 7, 2 crash on altsun1 and 2 on vlsi06. With 8 no crashes at all! Afterwards RPC timeout has been increased up to 200 and there have been no crashes up to 14 jobs per client.

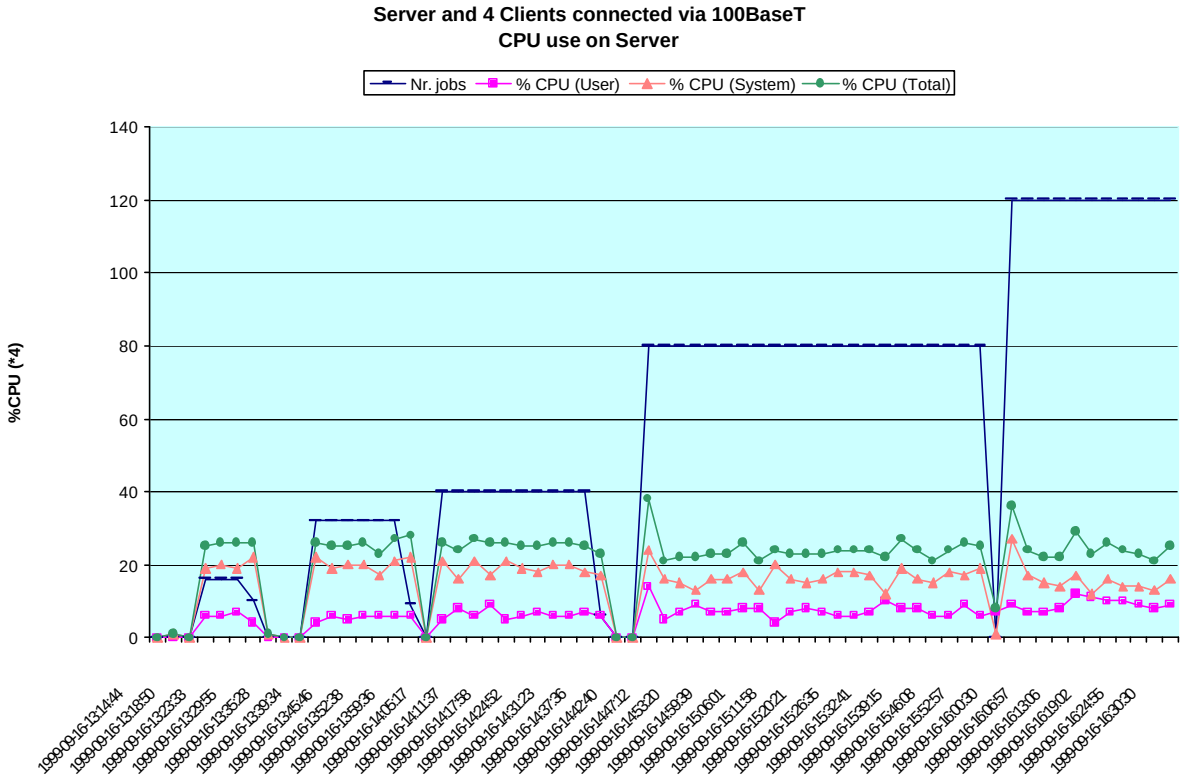




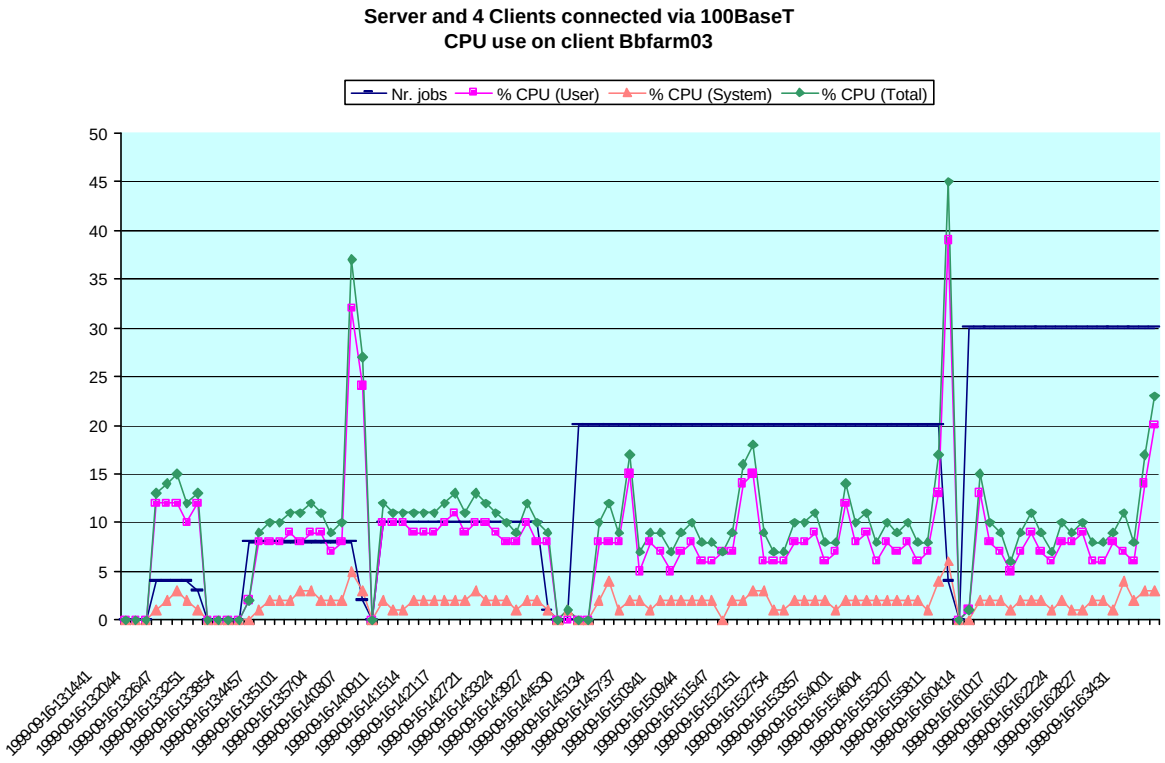
It appears that the mean wall time is influenced by the link speed: higher is the line speed and smaller is the MWT of the job. Moreover the job MWT on gsun client is high in comparison with the correspondent value in LAN Test 1(with 6 jobs MWT is 1200sec versus 200 sec of Test 1) and this has to be investigated.

4.2 LAN 100Mbps Test

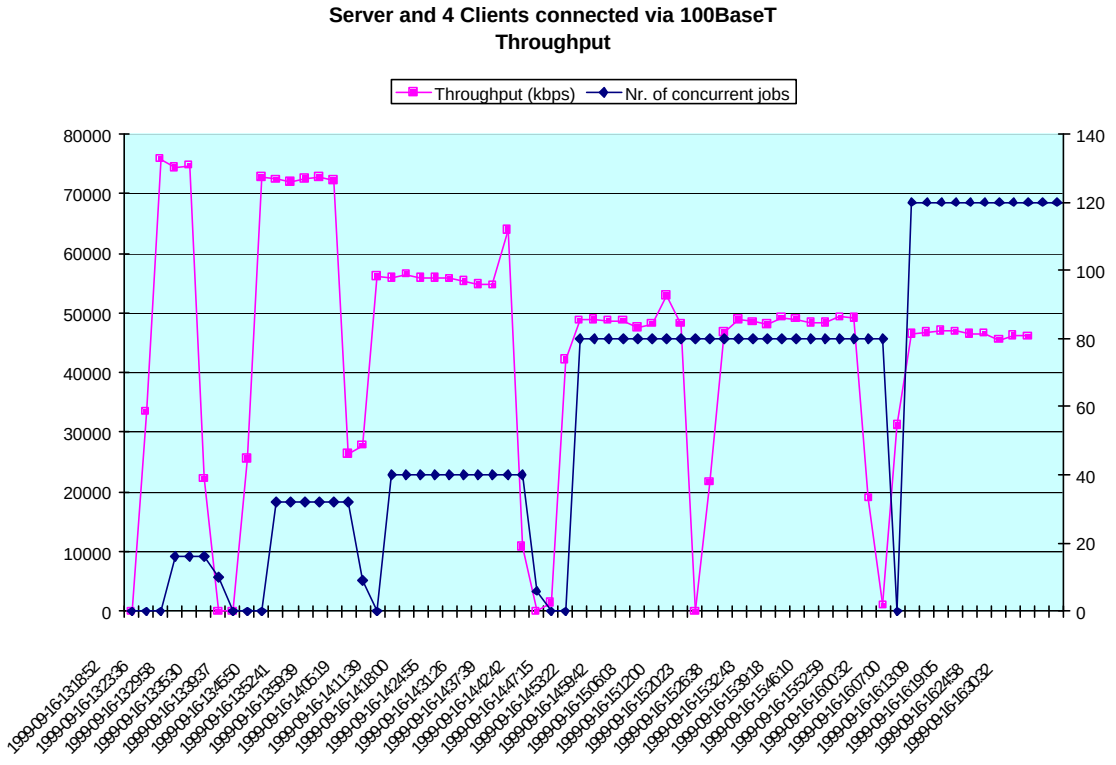




Server CPU is saturated with 18 jobs, as in Test 2.



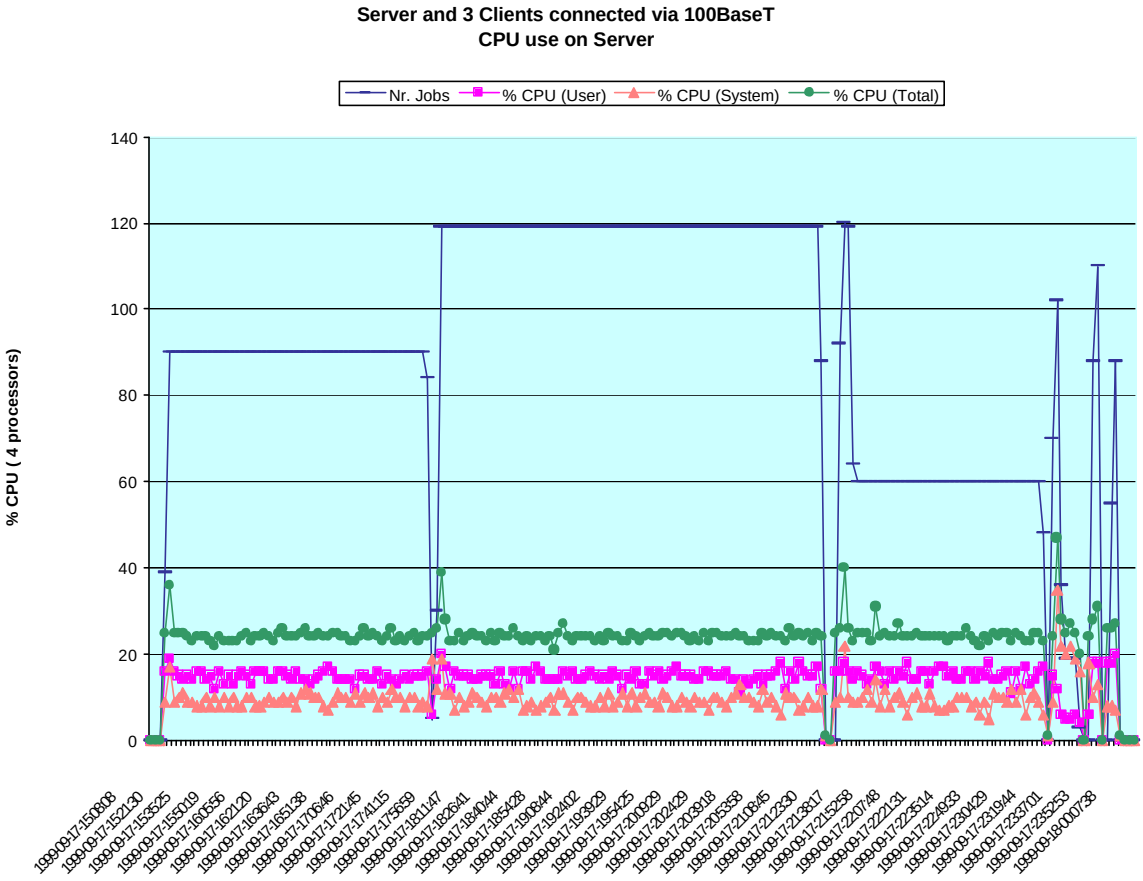
Being a 4 CPU system, this client works far from CPU saturation.

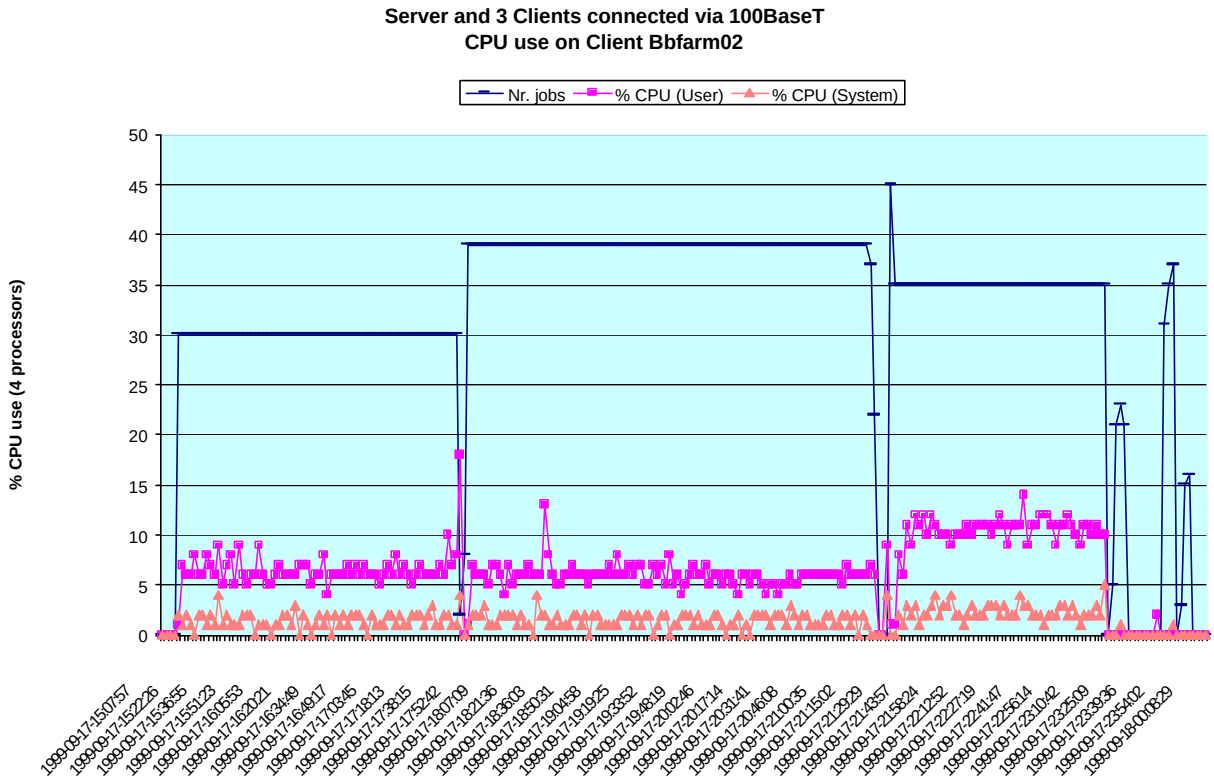


Throughput is good as long as server CPU use is not 100 %.

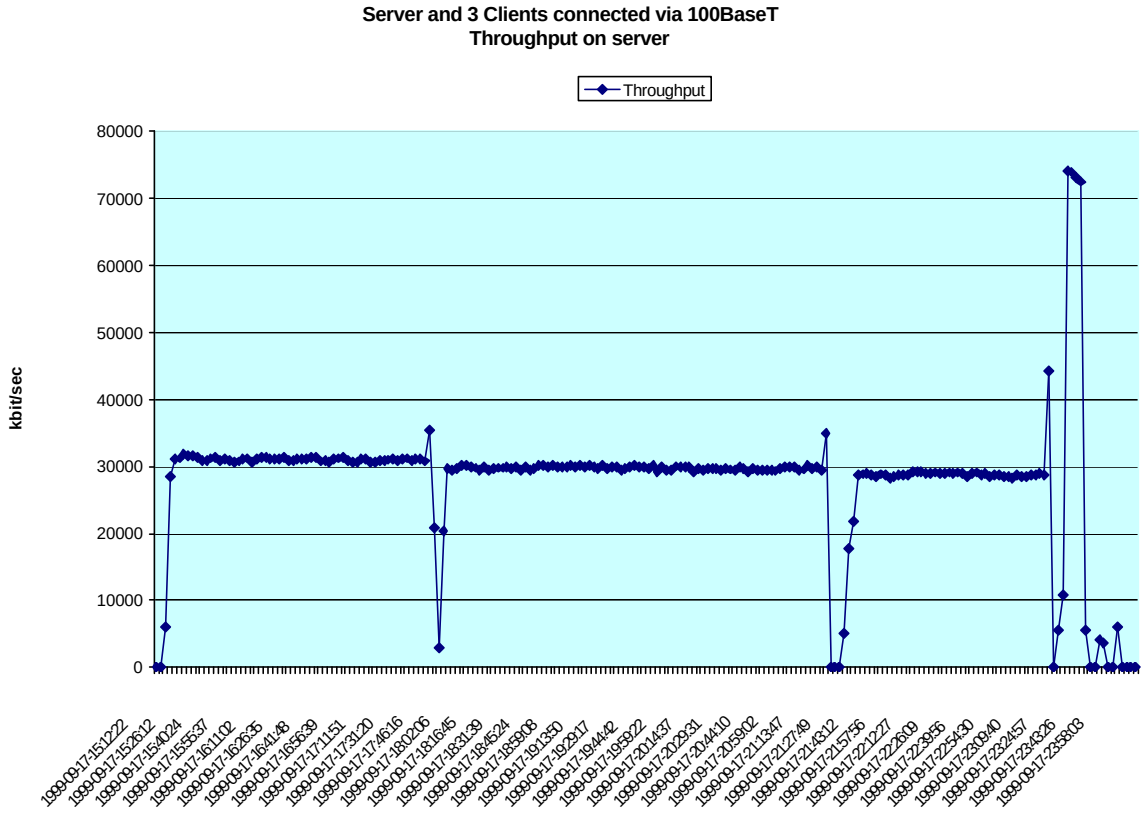
This figure confirms that the system works well below 20 jobs.

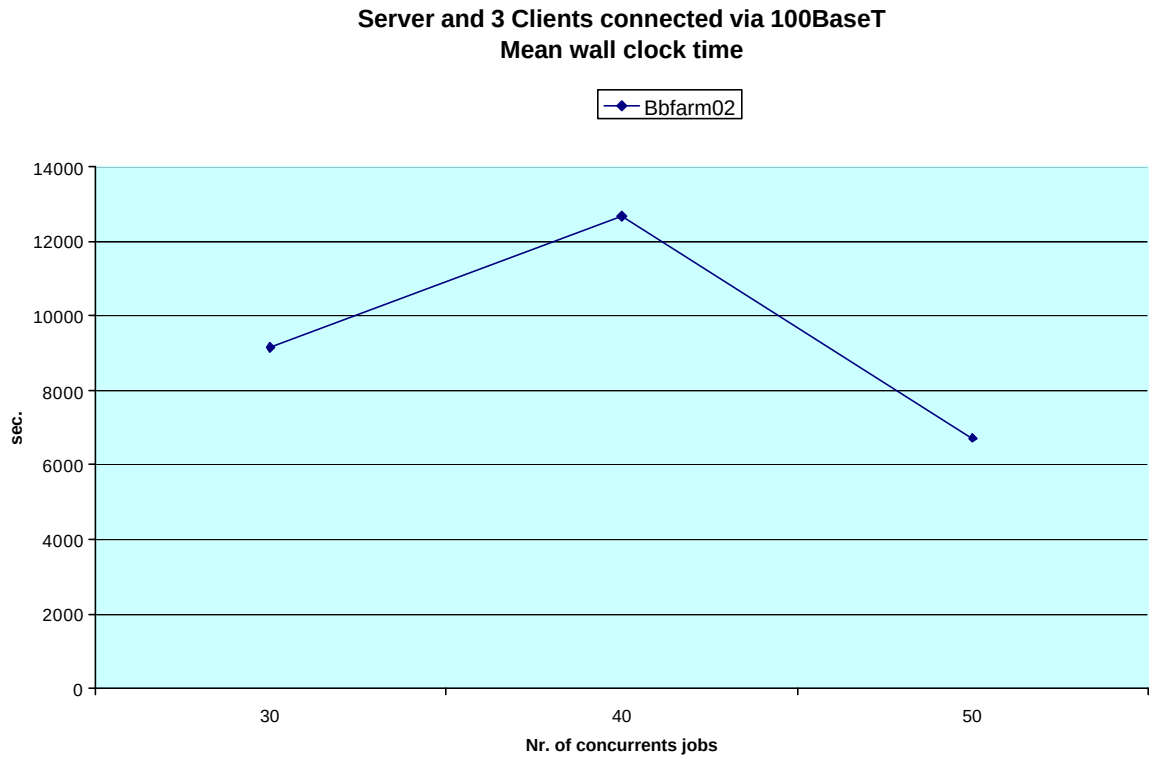
Figures starting from 30 concurrent jobs up to 45 per clients and with 3 clients follow.


Server and 3 Clients connected via 100BaseT
CPU use on Server
Nr. Jobs
% CPU (User)
% CPU (System)
% CPU (Total)
% CPU (4 processors)
140
120
100
80
60
40
20
0
1999-09-17-150808
1999-09-17-152130
1999-09-17-153525
1999-09-17-155019
1999-09-17-160556
1999-09-17-162120
1999-09-17-163643
1999-09-17-165138
1999-09-17-170646
1999-09-17-172145
1999-09-17-174115
1999-09-17-175659
1999-09-17-181147
1999-09-17-182641
1999-09-17-184044
1999-09-17-185428
1999-09-17-190844
1999-09-17-192402
1999-09-17-193929
1999-09-17-195425
1999-09-17-200929
1999-09-17-202429
1999-09-17-203918
1999-09-17-205358
1999-09-17-210845
1999-09-17-212330
1999-09-17-213817
1999-09-17-215258
1999-09-17-220748
1999-09-17-222131
1999-09-17-223514
1999-09-17-224933
1999-09-17-230429
1999-09-17-231944
1999-09-17-233701
1999-09-17-235253
1999-09-18-000738



After 40 concurrent jobs on client, jobs start crashing due to timeout with server (>120 job) and throughput increases again.





It could be interesting to notice that when jobs start crashing and throughput increases up to the maximum, mean wall clock time values decreases

5 Conclusions

Tests provide a description of Objectivity configuration on different network layouts, with different link characteristics, in terms of CPU behavior, link throughput and job execution time measures. Sun single and multiprocessor systems have been used.

The inability of Objectivity AMS 5.1 to use multiprocessor systems represents a bottleneck with network speed at 100 Mbps. The high CPU usage also on Sun multiprocessor clients running over 100 Mbps network enhances that Objectivity implementation is heavy and it could be improved. An important parameter of the different tested configurations is the number of connections on the server and the best measured value corresponds to 30 jobs, that is too small.

As exercise, analyzing the results it is possible to identify some boundary conditions for an efficient running of the jobs, with the CPU powers used. Let us suppose that an 'efficient running of the job' is when mean wall clock time is less than 10 x job mean wall clock time, for each configuration the following boundary conditions sort out:

1. 1000 Mbps: 1 job=60 sec, 15 job=600, server CPU < 60%, client CPU ~100 % , throughput 37 Mbps

2. 100 Mbps: 1 job=48 sec (155 sec*), 15 job = 600, server CPU < 100%, client CPU ~50% , throughput 80 Mbps

3. 10Mbps: 1 job=200 sec, 6 jobs = 600, server CPU < 20%, client CPU ~50% , throughput 8.5 Mbps

4. 2 Mbps WAN Test 4: 1 job=1000, not acceptable

5. Wan Test 5(n×client): min wall clock time is acceptable only on Gsun and Atlsun1(8 Mbps)

6. 100 Mbps Test 6 (n×clients): 1 job 155, 6 jobs = 600, server CPU ~ 100%, client CPU ~ 15%, throughput 70 Mbps

* in this case the job reads 400 MB instead of 120 MB

On the basis of the above conditions, possible WAN scenarios should be based on link with a minimum speed of 8 Mbps between client and server. Client machines should run from 6 up to 15 concurrent jobs and server should deal with requests of 30 concurrent jobs as a maximum.

5.1 Future work

Testing Objectivity 5.2 has the highest priority. Since it is supposed to be 'multithreaded' and therefore able to use multiprocessor systems in efficient way, we should be able to actually test the scenario with 4 SUN E450 connected at 100 Mbps and then at 1000 Mbps.

An other important test will be to configure one server and several clients on WAN with link at 10 Mbps in order to compare the results with those obtained with the same speeds on LAN. It will be important to check if delays at that speed will affect the results.

Multi-server configurations using read/write applications will be taken into account for future tests.

Since 100 Mbps seems to be a reasonable speed and it allows good job wall clock time, WAN layout at 100 Mbps would be very interesting for testing; the configuration could be: 3 servers (multiprocessor), 50-100 clients per server, 10-20 jobs per client (depending on the CPU power and architecture).

A WAN layout at 100 Mbps is quite reasonable if we think that the present European research networks are based on 155 Mbps backbones and many of them are planning backbones based on Wave Division Multiplexing links that are able to provide n x 2.5 Gbps (where n can range from 8 up to 100). The European backbone, TEN-155, is planning to evolve to a new gigabit backbone (Geant project). Time scale for Gigabit backbones is 2000-2001 therefore network access link of 100 Mbps for client machines should be easy to get.

Acknowledges

We would like to thank Francesco Prelz for his contribution to the tools development and Emanuele Leonardi, who made available Babar farm for our tests.

References

[1] http://atlasinfo.cern.ch/Atlas/GROUPS/PHYSICS/HIGGS/Atlfast.html

[2] M. Boschini, L.Perini, F. Prelz, S. Resconi "*Preliminary Objectivity tests for MONARC project on a local federated database*", MONARC Note 99/4