PROPOSAL

HEP Batch Computing Facility based on a PC Farm

Corrections:
November 3, 1997
November 28, 1997

A. Chevel
shevel@pnpi.spb.ru

Creation date: August 20, 1997

Abstract

This proposal is devoted to realization and practical use a General Purpose Batch Computing Facility based on a PC farm for High Energy Physics research.

For up to date information please see "http://www.pnpi.spb.ru/pcfarm/".

Problem: Needs for high performance computing for a low cost per unit of power.

Introduction

Several considerable facts might be taken into account:

- decreasing the financial resources in HEP;
- increasing the requirements for computing;
- existing of a cheap operating platform like Linux;
- cheap commodity hardware (PC);
- existing batch systems CODINE, NQS, LSF, and others which are relatively easy available in many leading HEP laboratories;
- almost all important applications in HEP might be running on Linux;
- new languages: Java;

We might consider General Purpose Computing Facility on a base of PCs.

It is discussed PC farm consisting of N nodes (PCs) for batch processing with relatively weak communications between nodes. N might be considered in range from several PCs to probably 100 or more. There is already known computing system based on 9K (!) processors PentiumPro200 (Intel).

Reliability of an existing hardware is enough to support PC farm consisting of 100 nodes by one person (for hardware, operating system and general application software).

Scalability: we can increase a number of PCs to desired value in step by step mode by small quanta.

The value about **400 Pentium Pro 200 MHz or 200 Pentium II 300 MHz** can be discussed as really achievable maximum

of computing power (in case when four processor motherboards will be used).

One processor Pentium II 300 MHz is more than two times more powerful the R4400 200 MHz. That means the maximum computing power will be more than 30 SGI with 28 R4400 processors every SGI. In other words Pentium II 300 MHz has about 11 SpecInt95 and about 8 SpecFp95 (about 100 MFlops in Linpack).

A number of SCSI ports and disk drives may be various and will reflect concrete needs in the discussed proposal.

Technical consideration

We can consider large installation like a constellation of simple farms (see Fig. 2).

However before realization of the large farm we should create pilot relatively small experimental Batch Computing Farm (see Fig. 1).

There are several technical questions which may be mentioned now:

- how to maintain and load one OS version in every computer;
- how to use one (or two) keyboard and monitor for a hundred computers;
- how to display the status of the whole large farm in compact and consistent manner.

It is possible to continue this list however those technical details will be solved later because good part of them are already solved in a range of previous projects (FNAL: computing farm, DESY: Hermes PC Farm, etc.). We also have to take into account that new equipment, software, prices are appeared every month.

Experimental Batch Farm

To test scalability and other parameters discussed proposal it required to deploy small Batch PC Farm consisting of at least 3 PCs.

Main aim: to get real experience in deploying the batch system (CODINE,LSF, etc) on small PC Farm and to collect user feedback, measurement the real farm parameters: comparing the execution time for selected jobs. It is planned such a farm will be used for simulation and microDST analysis.

In any case this pilot project will show for which type of calculations this farm will fits more.

To deploy such an experimental installation we need for things which should be bought or found out:

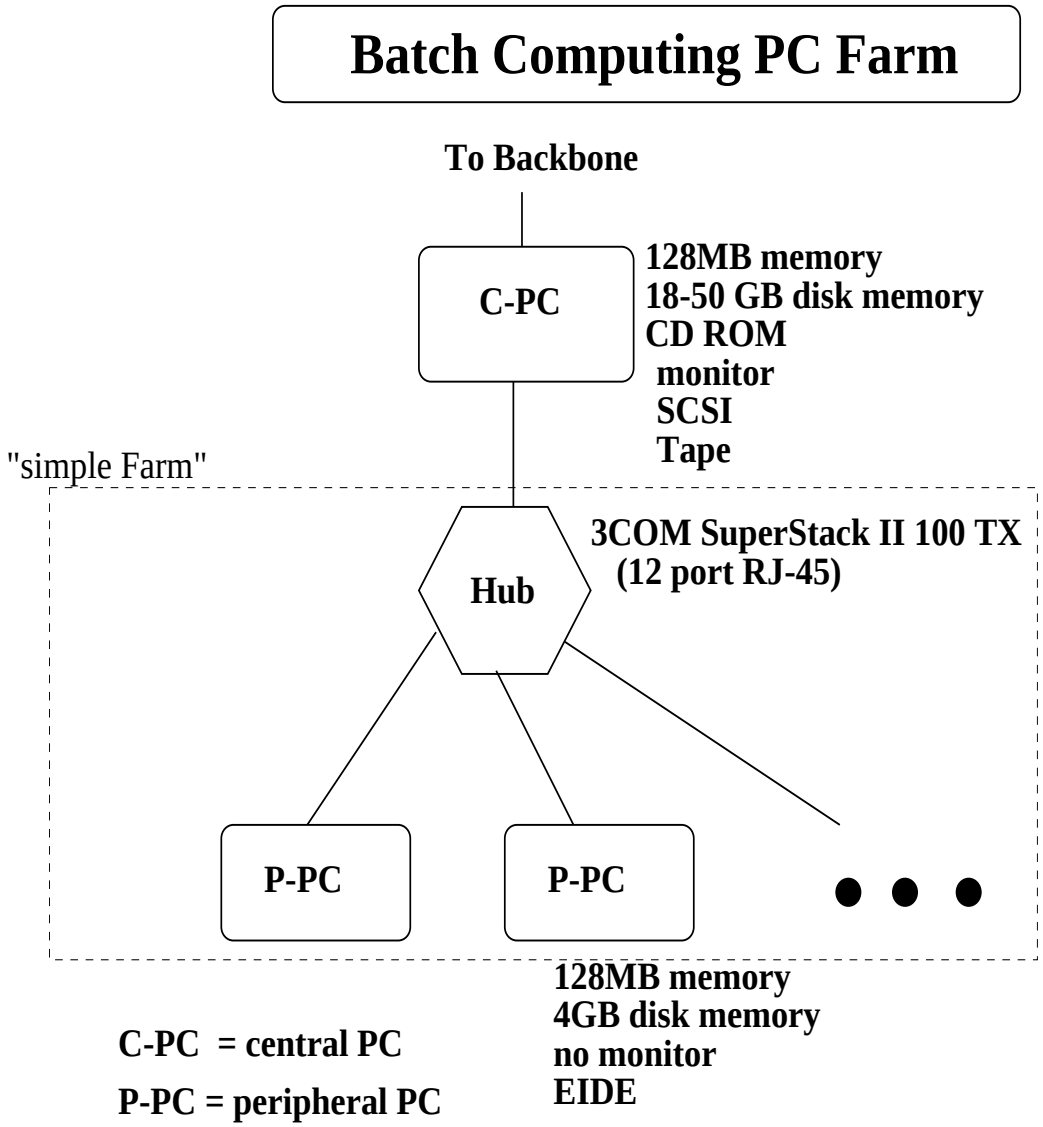
- two PCs:
  - Pentium II 300 MHz;
  - 512K cache;
  - 128MB main memory;
  - 4 GB EIDE disk drive;
  - video card (on mother board);
  - no monitor;
  - Fast Ethernet card;
  - total price for one PC is about $2700;
- Hub Fast Ethernet: 3COM Super-Stack II 100 TX (12 port RJ-45) - about $1100;
- APC Smart-UPS 700 - about $400;
- Linux SUSE distribution;
- batch system (LSF,CODINE,etc);
- central PC:

  - Pentium II 300 MHz;
  - 512K cache;
  - 128MB main memory;
  - SCSI controller (wide);
  - two disk drives (4 GB + 27 GB) ;
  - video card at least 4 MB;
  - two Fast Ethernet card;
  - CD ROM;
  - tape (?);
  - monitor 17";
  - total price for central PC is about $6K;

- total price for experimental installation is about $12900 for the start installation.

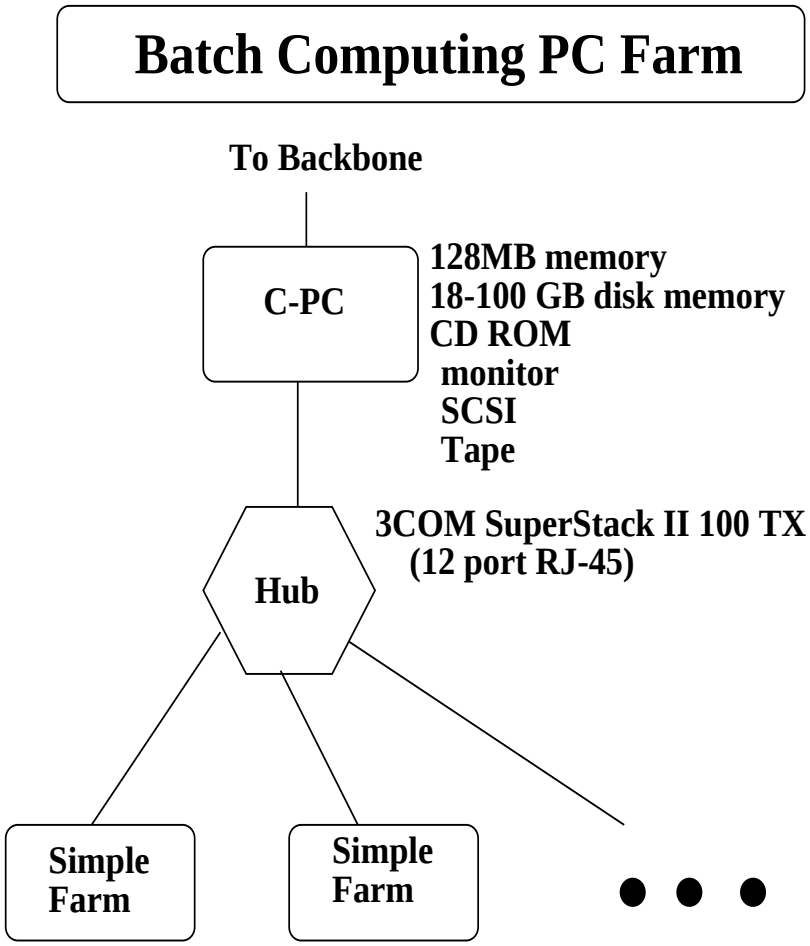
Things which should be done:

- hardware ordering and deploying;
- software installation and tuning:
  - Linux;
  - CERN Program Library;
  - gnu stuff and other standard de facto utilities;
  - www page(s) which will reflect the work progress;
  - batch software;
- measurement of execution time for the selected jobs and other parameters;
- organizing the user access (over Internet as well).



a) experimental Farm

Fig. 1. Experimental Batch Computing PC Farm.



C-PC = central PC

b) Batch Computing Farm

Fig. 2. A Scheme of the Batch Computing PC Farm.

Schedule:

Week	Job
0	Project start
4	Hardware delivery at PNPI
6	Linux + CERNlib + other programs, libraries, utilities
8	Start User access
10	Report paper about first results

Spending:

What	Who
Manpower	PNPI
space to deploy	PNPI
equipment	?
batch system	?
Linux SUSE	?

Final remarks

A tape cartridge may be used to transport the data to discussed computing facility from other places where such data are produced. It is not clear yet (27 Oct 1997) which tape drive type will be more suitable for that. Seemingly most useful to have the same cartridge type (and density) which is used in tape robot at CERN DESY, FNAL, or BNL.

For the pilot installation uniprocessor configuration is discussed. However it is easy to image how to upgrade it to two or four processor configuration.

Remarks after November 27, 1997

After exchange of opinions with several colleagues we began to think about DLT tape drive (not less 20 GB uncompressed) as a data transfer

media.

We began to think about dual processor configuration + 256 MB + 3*9GB disk drives on the central PC.

We take also into account that we will have some additional spending in accompanied accessories for the farm: shelfs, connectors. Some of the parameters might be changed (number of processors, memory, network equipment: e.g. network switch instead network hub). Maximum total spending we can image for start configuration is about \$25K-\$28K (about \$8K tape drive, \$2K shelfs for PCs, \$3K for switch, etc.)

HEP Division at PNPI approved such a pilot project in general.

Also we had several suggestions from computer companies to help us in various areas.

Incomplete list of the people who shown interest to this proposal

- Michael Ernst (DESY)
- Harry Renshall (CERN)
- Valery Schegelski (PNPI)
- Nikolai Terentiev (PNPI)
- Oleg Prokofiev (PNPI)
- Alexei Vorobyov (PNPI)
- Alexander Kiselev (PNPI)
- Sergei Kruglov (PNPI)
- Vladimir Korenkov (Dubna,JINR)
- Nikolai Kuropatkin (PNPI)
- Vladimir Abaev (PNPI)
- Wolfgang Wander (MIT)

Now we have more than 200 hosts around the World from which people investigated proposal www pages
"http://www.pnpi.spb.ru/pcfarm".