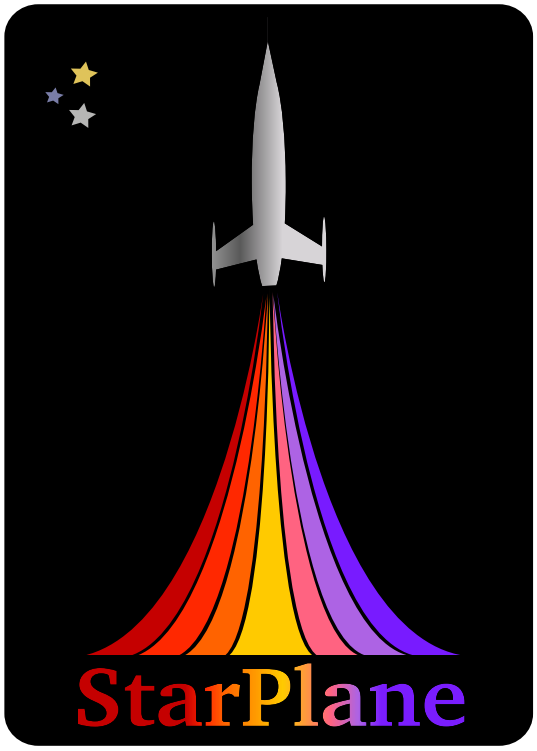

Application Use Cases

Jason Maassen
Version 1

vrije Universiteit amsterdam
UvA Universiteit van Amsterdam
NWO
SURFnet

Contents

1 Introduction

This documement contains a description of application use cases for the starplane project. For each usecase, a generalized description is given, followed by more detailed examples. Each usecase also lists the communication patterns used, the amount of data transfered, and the amount of time allowed for network reservation.

2 Usecase 1: Large 'stand-alone' file transfers

In this usecase, the optical network is used for large file transfers. These are 'stand-alone' file transfers which are performed directly, by a user or application and are not part of a scheduled job. Therefore, the transfers are typically performed between the frontend machines.
The bandwidth requirements in this usecase are 'soft'. In general, more bandwidth is better, but the storage hardware may limit determine the upper limit. Also, the user/application is often willing to settle for the amount of bandwidth that is available.

Typical Data Size	:	up to several TB
Typical Communication pattern	:	point-to-point, star (in)
Allowed Reservation Time	:	seconds to hours
Required Network Information	:	bandwidth available a certain point in time

2.1 Simple User-Driven File Transfers.

Although many DAS users stick to their local cluster, there are several power users that use multiple clusters. Since the systems of the clusters are not shared, these user regularly transfer files between sites, e.g., to synchronize their home directories using scp. Although the size of these transfers is generally limited to a couple of GB, they also tend to be 'interactive', where the user is waiting for the result. Therefore, reducing the transfer time from several minutes to several seconds may be a big improvement from a users point of view.

Typical Data Size	:	up to several GB
Typical Communication pattern	:	point-to-point
Allowed Reservation Time	:	seconds
Required Network Information	:	bandwidth currently available

2.2 Nightly backup.

The user data on the DAS-3 clusters needs to be backed up regularly. This can either be done locally or in a more centralized fashion. One option is to have a large storage unit in one location (e.g., at SARA), and store the backup of the DAS-3 machines there. In this scenario it is possible to use the optical network to transfer the data files to this central storage.
The backup process typically runs once a night, and may have to backup up to 1-2 TB per site. The backup may run in parallel on all sites, provided that the central backup unit can handle the data rate. At 10 GBit/sec., it should take about 15 minutes to backup 1 TB of data.

Typical Data Size	:	up to 1-2 TB per site (many files)
Typical Communication pattern	:	star (in) with backup site in center
Allowed Reservation Time	:	hours, runs at fixed time at night
Required Network Information	:	bandwidth available when the backup is run

2.3 Transfer of Large Medical Data Files (MRI scans)

Several large hospitals in the Netherlands share data files (MRI) for medical research. Currently these files are physically transfered using a harddisk, a box, and a courier service. Due to the limited size of harddisks, the total data size is smaller than 500G. The transfer does not have a very strict time limit (since the latency of a courier service is relatively high), but there are security issues, since the disk contains medical data.
Altough this example can not directly benefit from the DAS-3/StarPlane infrasturcture, we can come up with a 'proof-of-concept' which shows that large files containing sensitive data can be transfered quickly and securely using lightpaths.

Typical Data Size	:	up to 500 GB
Typical Communication pattern	:	point-to-point
Allowed Reservation Time	:	hours
Required Network Information	:	when is bandwidth available
Extra Requirement	:	secure transfer

3 Usecase 2: Large File Stage-in/Stage-out.

In this usecase the optical network is used by the scheduler to transfer large input and/or output data sets to the machines selected to run a (parallel) application. The application itself has modest communication requirements and does not use the optical network. It simply requires the (large) input data set to be present on startup and/or produces a large output data set which must be returned to a location specified by the user. The scheduler is responsible for the bandwidth reservations and file transfer. The application has no knowledge of the StarPlane network.
The Koala scheduler (developed in Delft [3]) already handles file stage-in in a 'bandwidth aware' way; it schedules jobs as close to the data as possible. By providing up-to-date information on available bandwidth and network configuration times to Koala, it can decide which combination of site (job waiting time) and bandwidth yields the fastest startup time. It can then perform the required network reservation, transfer the data files, cancel the reservation, and starts the job.
The bandwidth requirement in this scenarion are 'soft'. For example, 10 Gbit/s to an empty site may result in a faster startup time then 40 Gbit/s to a busy site. The disk performance may also limit the transfer speed.
The software infrastructure required for this usecase is very similar to that used in Usecase 1. The main difference is that the scheduler may require available bandwidth information over a longer period of time.
This usecase illustrates how the StarPlane infrastructure can be used to increases the utilization of DAS-3, since it gives users with data intensive applications more options on where to run their applications.

Typical Data Size	:	up to a TB
Typical Communication pattern	:	point-to-point, star (out)
Allowed Reservation Time	:	minutes to hours
Required Network Information	:	available bandwidth in future (series)

3.1 MEG modelling.

The Vrije Universiteit Medical Centre (VUmc) posesses a Magnetoencephalography (MEG) scanner which produces large amounts of data which must be processed. MEG is a tool to study the function of the human brain. It measures the magnetic field intensity at hundreds of points over the surface of the skull up to several thousand times per second. Measurements made while a subject is recognizing a picture, performing mathematical calculations, watching an alternating check board pattern, sitting quietly with their eyes closed, or a host of other tasks, provide insight into the functioning of the brain, whether healthy or diseased.
The size of a data set from one session with one subject is typically hundreds of megabytes. When this is multiplied by multiple sessions per subject and dozens of subjects in a study, the computational demands become arduous and clustered computer resources are necessary.
The current version of the application is run in a task-farming fashion and the individual processes do not communicate. Each process expects the data files to be available on the local hard disk of a node, since this makes reading the data much quicker and easier. Plans exist for more advanced versions of the application which require significantly larger amounts of data.

Typical Data Size	:	many files up to 1 GB each
Typical Communication pattern	:	point-to-point, star (out), scatter within cluster
Allowed Reservation Time	:	hours, application typically runs at night
Required Network Information	:	available bandwidth at start time indicated by user

3.2 Analysis of Video Data.

MultiMedian [4] has developed software capable of performing parallel video anaylsis of large amounts of video data. This software is used to extract a wide range of 'features' from the stream, which can be used to search in the video data. This software has been used succesfully in the Trecvid competition [1], There is also some interest from the Beeld en Geluid Institute [5] in Hilversum, who would like to use this technology to create a 'search-engine' for the Dutch video archive. This archive contains approxemately 750.000 hours of (mostly analog) audio and video recordings. Each year some 10.000 hours of new high-definition digital video recordings are added and 10.000 hours of old analog recordings are converted to low-definition digital.

Typical Data Size	:	250 GB to several PB
Typical Communication pattern	:	star (out), with MultiMedian or Beeld en Geluid at the center
Allowed Reservation Time	:	hours, typically runs as off-line processing
Required Network Information	:	available bandwidth in future (series)

4 Usecase 3: Applications with Static Bandwidth Requirements.

This usecase consists of (parallel) applications which have high communication requirements, possibly in addition to large in and/or output files. The communication requirements do not change during the lifetime of the application.

Typical Data Rate	:	up to several tens of GBit/sec
Typical Communication pattern	:	star, all-to-all
Allowed Reservation Time	:	minutes to hours
Required Network Information	:	available bandwidth in future (series)

4.1 Distributed Game-Tree Search.

Many game playing programs work by analyzing millions of positions that could arise in the next few moves of the game. Algorithms for searching positions in parallel often rely on distributing the search space. Work is then migrated from the machine that generated it to the machine which 'owns' the correct part of the search space. As a result, these algorithm communicate heavily and randomly.
In 2002 John Romein solved the game of Awari using retrogade analysis. The entire DAS-2 cluster at the VU (144 processors) solved the game in 51 hours. The resulting position database is 778 GB. Almost 900 billions states where searched producing 130 TB of communication. Although the amount of communication per processor was not extremely high (a maximum of 120 MBit/sec), the peak communication thoughput of the entire cluster was some 17 Gbit/s.
At the time, the performance of the wide-area links of the DAS-2 was far from sufficient to run the application on multiple clusters. The combination of DAS-3 and StarPlane should be capable of this, however. If we assume the application was computation or memory bound (which is likely, given the resonably low communication rate per processor), we can expect it to run at least twice as fast on a DAS-3. processor. This would increase the datarate to approximately 240 MBit/sec per processor, or 34 GBit/s for a 144 processors run. Increasing the number of machines and distributing them over multiple clusters may result in a scenario which can use the full bandwidth of the optical network.
Note that it may be feasable to use the DAS-3/Starplane to solve a more complicated game then Awari, e.g., Othello.

Typical Data Rate	:	up to several tens of GBit/sec
Typical Communication pattern	:	all-to-all
Allowed Reservation Time	:	minutes to hours
Required Network Information	:	time range in which available bandwidth is sufficient

4.2 Remote Data Access for Analysis of Video Data

The user of the "Analysis of Video Data" example application has indicated that he's also interested in streaming data from a storage location instead of performing a stagein/stageout of the files. The application typically reads a couple of frames of video data, after which it spends a few seconds processing this data. When it is done, it reads the next couple of frames, etc. This would result in a scenario where a steady stream of data would be required between clusters. The data rate per machine would not be very high, but the total cluster may require several GBits/s in order for it to be to be usable. When multiple clusters are used the total data rate at the storage site may be significant.

Typical Data Rate	:	several GBit/s per participating cluster
Typical Communication pattern	:	star (out)
Allowed Reservation Time	:	minutes to hours
Required Network Information	:	time range in which available bandwidth is sufficient

4.3 Remote Visualization.

ASK CEES / ANDRE

5 Usecase 4: Applications with Dynamic Bandwidth Requirements.

This usecase consists of (parallel) applications which have high communication requirements, possibly in addition to large in and/or output files. The communication requirements change during the lifetime of the application.

Typical Data Rate	:	up to several tens of GBit/sec
Typical Communication pattern	:	star
Allowed Reservation Time	:	seconds to minutes
Required Network Information	:	?

5.1 Remote Data Access for MEG Modeling

The user of the "MEG Modeling" application has indicated that he is also interested in streaming data from a storage location instead of performing a stagein/stageout of the files. The application typically reads a large block of data, after which it spends a significant amount of time processing this data. When it is done, it reads the next block of data, etc. This leads to a scenario where a machine would require several gigabit/s for a couple of seconds, after which the network would be released for a longer period of time.

Typical Data Rate	:	peaks up to several GBit/sec
Typical Communication pattern	:	star (out)
Allowed Reservation Time	:	seconds to minutes
Required Network Information	:	?

6 Other ideas

This section contains some other ideas for example applications.

6.1 Application Migration.

When a data intensive application is running out of time on one site, machines may be available on another site. Instead of checkpointing to disk, the application data may be send directly to the replacement machines on another cluster.

Typical Data Rate	:	?
Typical Communication pattern	:	point-to-point
Allowed Reservation Time	:	seconds to minutes
Required Network Information	:	which site can provide the required bandwidth

6.2 Malleable / Adaptive Applications.

At the VU we have been experimenting with malleable/fault tolerant applications, which can handle changing numbers of processors during their lifetime. As long as a single processor remains, the application will keep running. We are currently extending this mechanism to support 'self adaptivity' (which allows an application to add or remove machines depending on how it judges its own efficiency) and checkpointing (which allows the application to survive even if the last machine if removed).

These mechanims could be combined into an application which is capable of maintaining itself. When it's time is running out on one site, it can reserve machines in a different site, migrate there, and continue to run.
Currently these mechanims are only capable of handling compute intensive applications, which do not use/produce large amounts of data. Maybe this can be extended to include the migration of large data sets ??

Typical Data Rate	:	?
Typical Communication pattern	:	point-to-point
Allowed Reservation Time	:	seconds to minutes
Required Network Information	:	which site can provide the required bandwidth

6.3 Analysis of Live Video Streams.

Video camera's are a commonly use to secure certain area's or buildings. When large area's need to be secured (e.g., an airport, highways, a city center), the number of camera's can become very large. So large, in fact, that it becomes hard for the security personel to actively monitor all video streams. In this case, automatic analysis of the video streams can provide a solution.
Each camera produces a continues data-stream of several Mbit/s which needs to be analysed and correlated with other video streams. As a result, the total data rate of all camera's in an airport can reach several GBit/s.

Typical Data Rate	:	several GBit/s
Typical Communication pattern	:	star (in)
Allowed Reservation Time	:	-
Required Network Information	:	-

6.4 Parallel File System.

IBM offers a "General Parallel File System" [2], which, simply put, uses the disks in a cluster as a software RAID system. Using this approach, they recenty managed to create 1.6 petabytes file system, with a 800 Gbit/s transfer rate. This experiment used a 1000-nodes of the ASC Purple supercomputer at Lawrence Livermore National Laboratory.
Although the DAS-3 will probably be a bit smaller than the ASCI Purple, the idea of using all disks in a cluster for one virtual file system is appealing. Assuming that the disks are capable of an 80 MByte/s transfer rate (640 Mbit/s), a virtual files system on a 32-node cluster should be capable of providing approximately 20 Gbit/s.

References

[1] Anual TrecVid Competition. http://www-nlpir.nist.gov/projects/trecvid.

[2] IBM General Parallel File System. http://www-03.ibm.com/servers/eserver/clusters/software/gpfs.html.

[3] H.H. Mohamed and D.H.J. Epema. Experiences with the KOALA Co-Allocating Scheduler in Multiclusters . In *Proc. of the 5th IEEE/ACM Int'l Symp. on Cluster Computing and the GRID (CCGrid2005)*, May 2005. http://www.st.ewi.tudelft.nl/koala/.

[4] MultiMedian. http://www.multimedian.nl.

[5] Nederlands Instituut voor Beeld en Geluid. http://www.beeldengeluid.nl.