

Putting Social Science Applications on the Grid

Rob Crouchley, Ties van Ark, John Pritchard, Dan Grose
Centre for e-Science, University of Lancaster

John Kewley, Rob Allan
e-Science Centre, CCLRC Daresbury Laboratory

Mark Hayes, Lorna Morris
Cambridge e-Science Centre, University of Cambridge

Introduction

- The need for HPC grid resources. An example - SABRE.
- The problems associated with using the grid.
- A solution. GROWL – an overview.
- Summary.

SABRE

Software for the Analysis of Binary Recurrent Events

SABRE is designed to model recurrent events for a collection of individuals or cases and many other types of repeated measures data with binary, ordinal or count responses. It fits both standard models and various mixture models which allow for residual heterogeneity.

It can be used to fit the following univariate statistical models :

- binary data with logit, probit or complementary log-log link
- ordinal response data using a probit link
- count response data using a log-linear Poisson model
- continuous response using identity link

SABRE employs reweighted least squares (standard homogenous models) and Newton-Raphson maximum likelihood (random effects binary models) algorithms. Both algorithms have been parallelised using MPI.

Run Time Comparisons. SABRE – Parallel SABRE - STATA

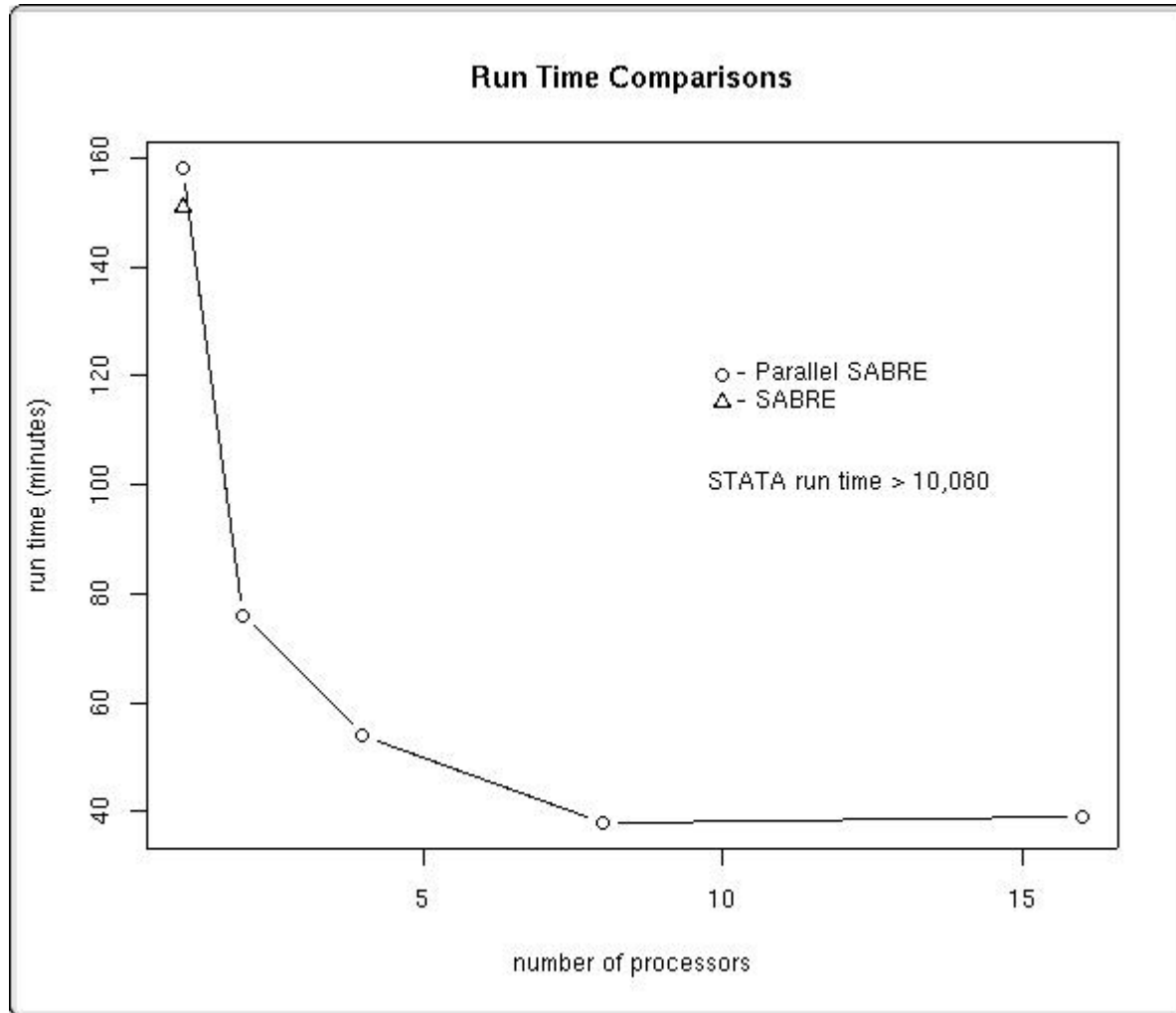
Comparisons :

- A random effects logit model fitted in STATA using the xtlogit command with 12-point quadrature.
- SABRE, logit link, 12-point quadrature
- Parallel SABRE, logit link, 12-point quadrature

Illustrative data, months to employment dataset contains 3,655,704 monthly observations on 199,881 individuals, with 14,716 non-zero binary outcomes. Comparison uses same starting values. The total number of model parameters = 54.

Run Time Comparisons. SABRE – Parallel SABRE - STATA

VRE



SABRE Developments

- Has been extended for bi-variate analysis.
- Will be extended to tri-variate analysis and greater.
- Computational time increases geometrically with the number of variates.
- These developments demand HPC resources.

The Problems

- Large number of software components required by client application to enable the Grid - e.g. security components, resource allocators, schedulers etc
- Components difficult to install and manage.
- No integration into existing client research applications (R, S, Stata, MATLAB etc.)
- No well defined 'work flows' - existing methods are 'ad hoc'
- ' ... making applications “Grid enabled” is seen by some as a distraction from getting real science done.' - J M Schopf & B Nitzberg. “Grids : The Top Ten Questions”

Classification of Grid User

(adapted from Foster and Kesselman)

Class of User	Purpose	Requires	Concerns
End users (e.g. quantitative social scientists).	Do research. Solve problems.	Applications.	Transparency, ease of use, performance.
Application developers.	Develop new and extend existing applications.	Programming tools, API's,libraries.	Ease of use, re-usability.
Tool developers.	Develop API's, toolkits, libraries.	Grid services.	Adaptivity, applicability, robustness, stability.
Grid developers.	Provide grid services.	Local system services.	Security, connectivity, protocols.
System administrators.	Manage grid resources.	System tools.	Balancing local and global concerns.

The Problems – A Missing Layer

Class of User	Purpose	Requires	Concerns
End users (e.g. quantitative social scientists).	Do research. Solve problems.	Applications.	Transparency, ease of use, performance.
Application developers.	Develop new and extend existing applications.	Programming tools, API's, libraries.	Ease of use, re-usability.
Tool developers.	Develop API's, toolkits, libraries.	Grid services.	Adaptivity, applicability, robustness, stability.
Grid developers.	Provide grid services.	Local system services.	Security, connectivity, protocols.
System administrators.	Manage grid resources.	System tools.	Balancing local and global concerns.

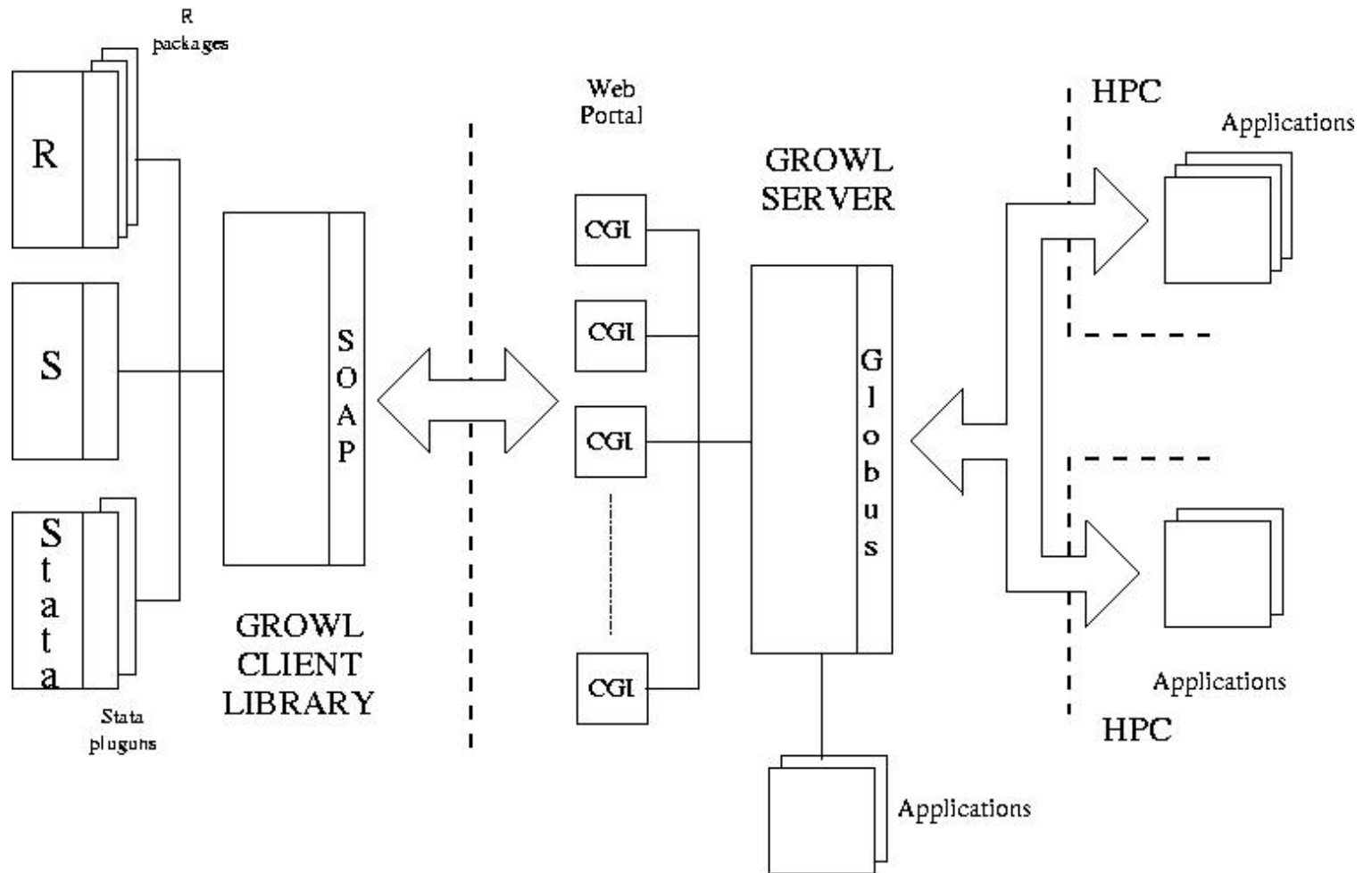
A Solution - GROWL

Grid Resources on a Work station Library

Project Objective – Demonstrate a lightweight client/server library that provides :

- Transparent client side handling of GRID related issues e.g security, file transfer etc.**
- Modules, libraries and “plug in's” that interface with existing client software tools.**
- Extensibility via a simple API with common language mappings (C++,C and Fortran).**
- A persistent multi-client server linked to existing grid components (primarily the Globus toolkit) providing access to HPC resources, session management, scheduling, authentication etc.**

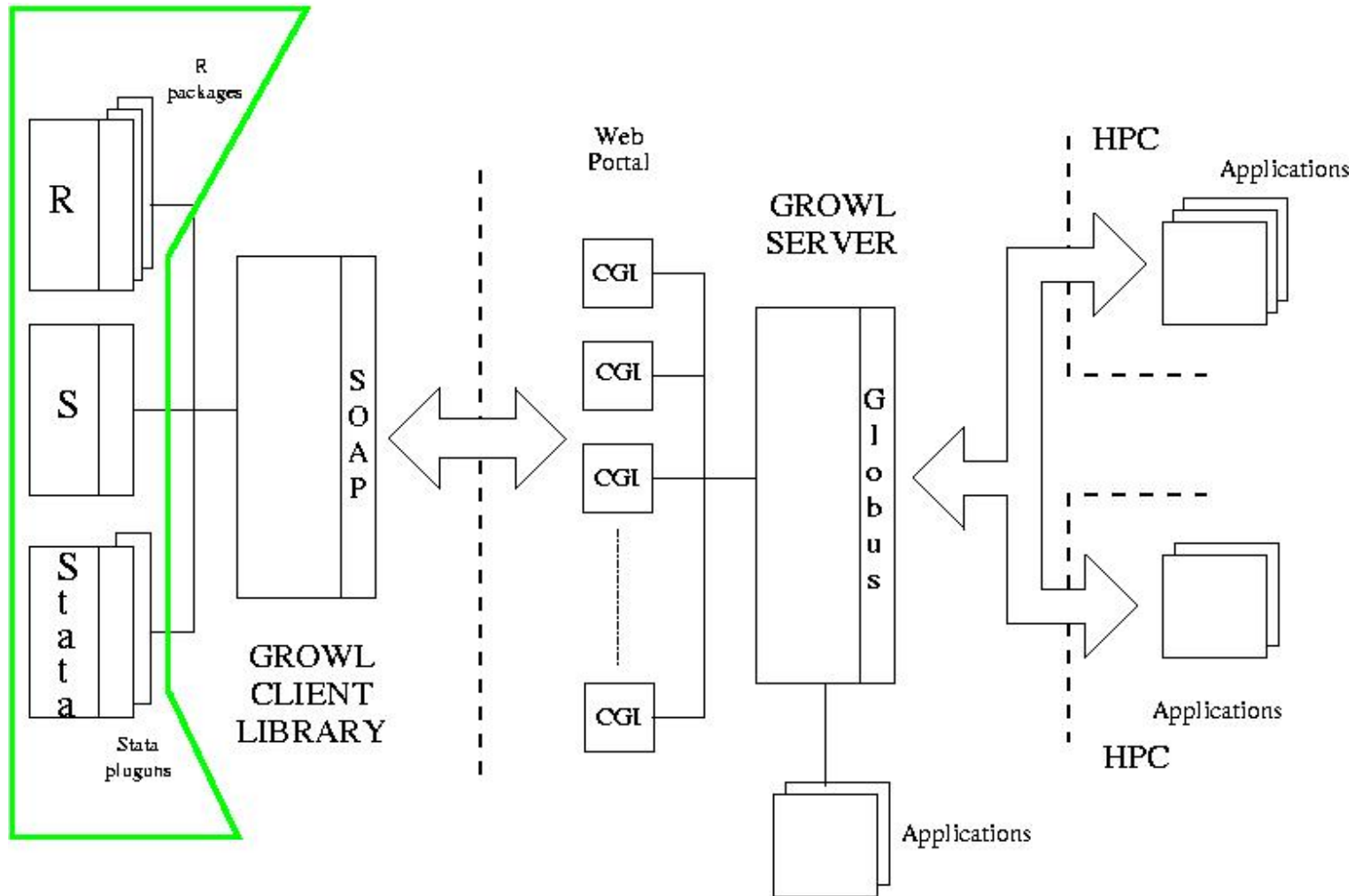
GROWL Architecture



GROWL Architecture

End User

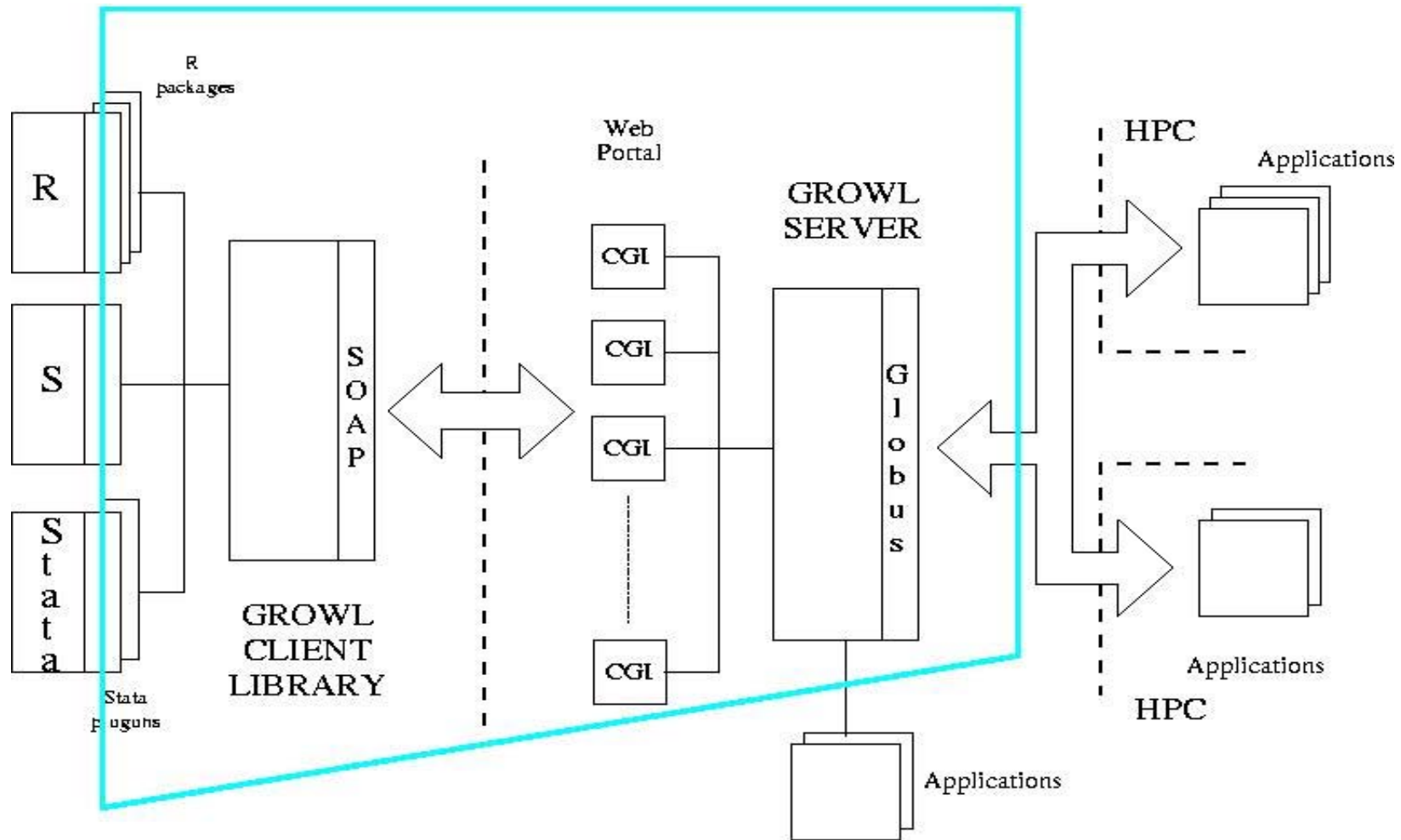
VRTE



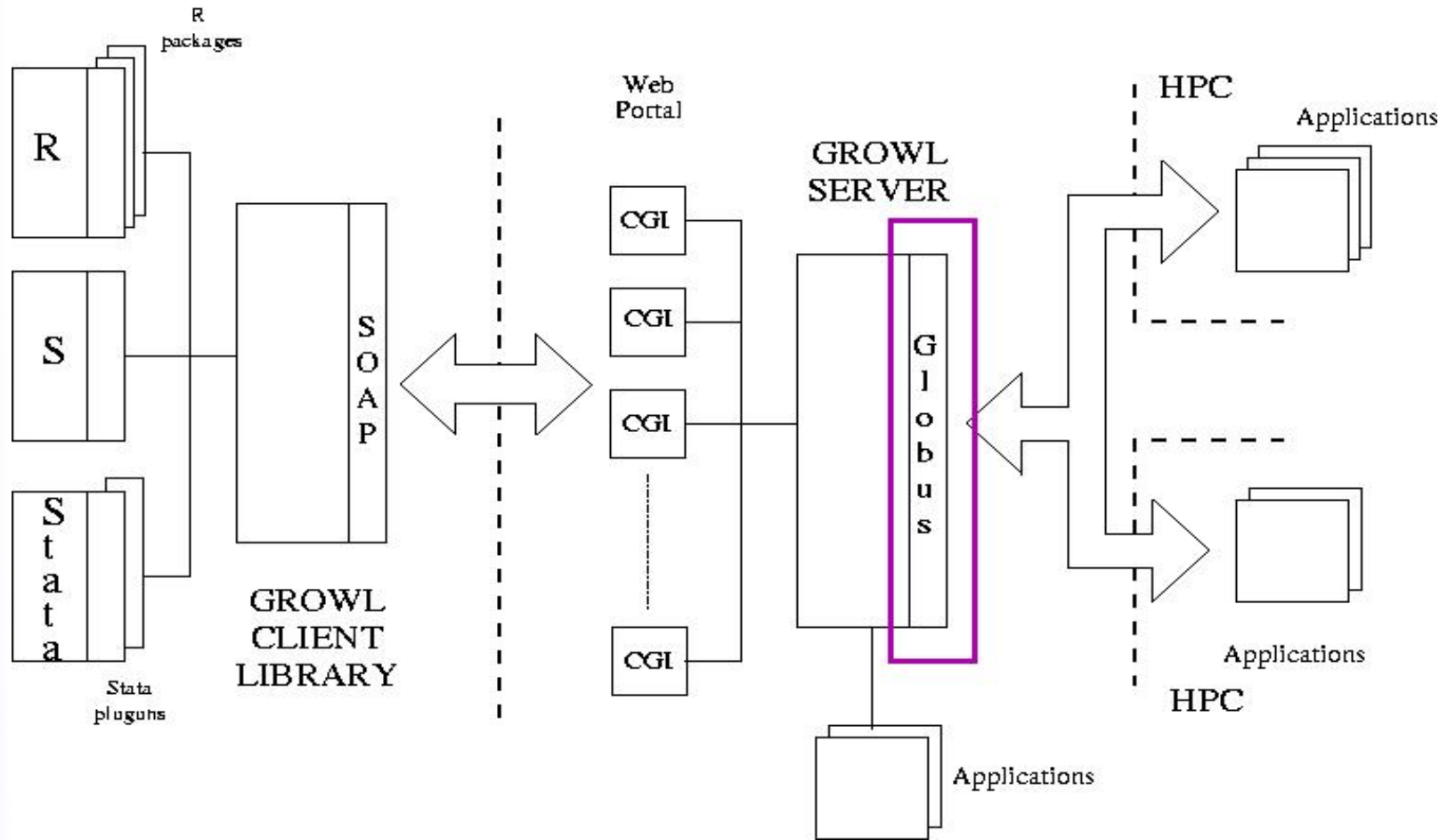
GROWL Architecture

GROWL Developer

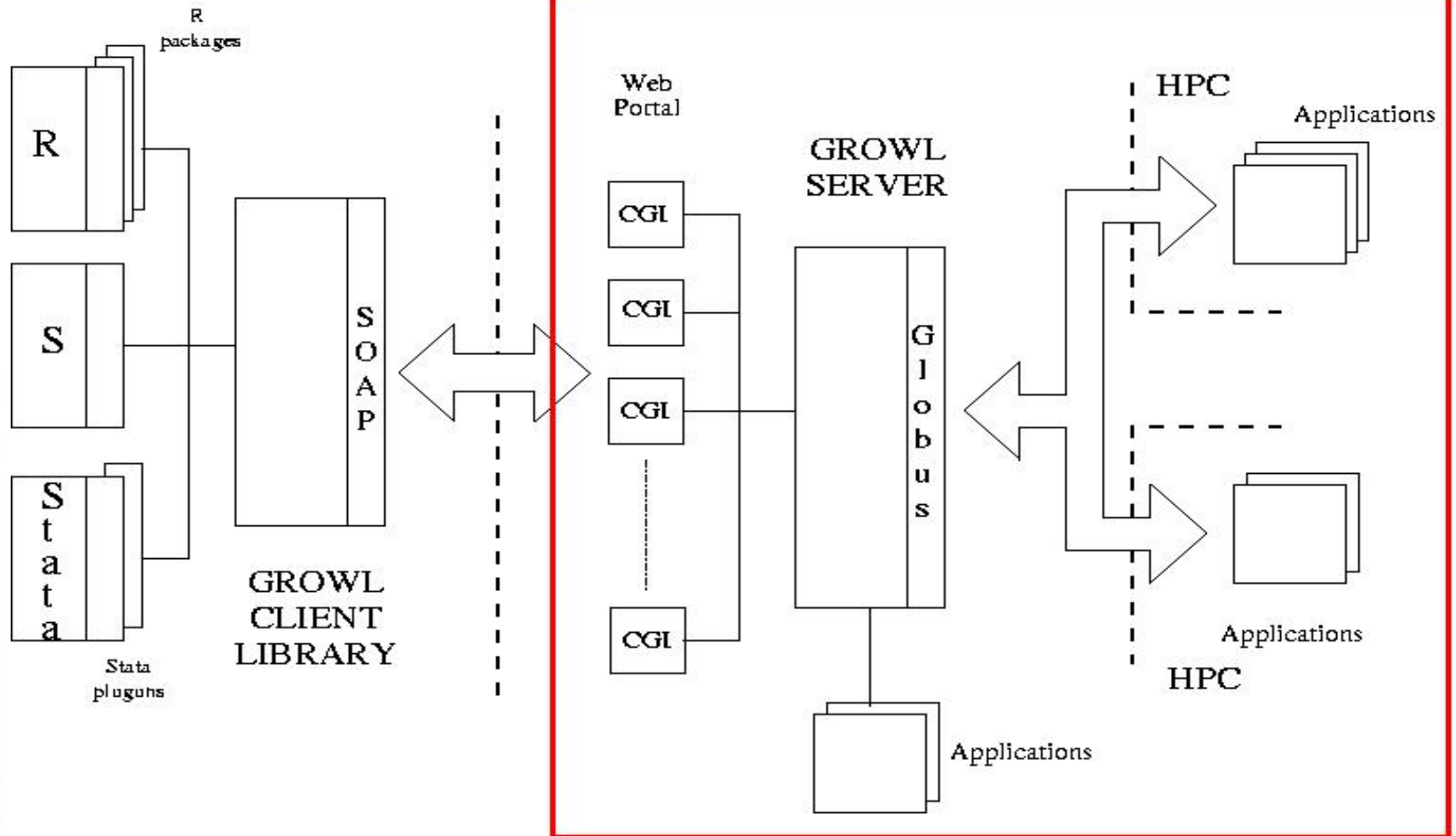
VRTE



GROWL Architecture Grid Developer



GROWL Architecture Systems Administrator



Sample R Session

"

```
> library("sabreR")
> sabre.session.0<-new.sabre.session()
> sabre.current.session(sabre.session.0)
> data<-list("pid","wave","hid","pno","ivfio","hoh","opfamb","opfamc","opfame",
+ "opfamf","opfamg","opfamh","sex","age12","race","region","jbsic","jbgold",
+ "jbstat","child","qfedhi","mastat",
+ "tenure","emp")
> sabre.data(sabre.session.0,data)
> sabre.yvariate("opfamb")
> sabre.ordinal(5,"cutpt")
> sabre.read("bhps.dat")
> variables<-list("cutpt","sex","jbgold","qfedhi","mastat","tenure","jbsic")
> factors<-list<("factor.cutpt","factor.sex","factor.jbgold","factor.qfedhi",
+ "factor.mastat","factor.tenure","factor.jbsic")
> sabre.factor(variables,factors)
> sabre.lfit(model="fcutpt",list("fsex",factors)
> sabre.fit()
> results<-sabre.results()
```

Summary

- GROWL provides scientists who have a limited interest in and knowledge of computing with an easy vehicle to improve their access to computational power.
- GROWL will make it feasible, cheap and efficient, to increase the size and complexity of models, the amount of data processed and to achieve results in a much reduced time frame.
- GROWL will provide tools to increase the effectiveness of large scale research and the efficiency of the researchers involved, while at the same time reducing costs by allowing continued use of existing often expensive applications by using plug-ins.
- GROWL will enable many quantitative social scientists to make the step to using e-Science technology to solve their problems.

References and Resources

“The Grid: Blueprint for a Future Computing Infrastructure.” I. Foster and C. Kesselmann (editors) 1998. Morgan Kaufmann Publishers.

“The Anatomy of the Grid – Enabling Scalable Virtual Organisations.” I. Foster, C. Kesselmann and S. Tuecke 2001. Intl J. Supercomputer Applications.

“Grids : The Top Ten Questions.” M. Schopf and B. Nitzberg.

GROWL URL: <http://www.growl.org.uk>

SABRE URL : <http://www.cas.lancs.ac.uk/software/sabre3.1/sabre.html>

JISC VRE URL: <http://www.jisc.ac.uk/>