EFDA–JET–CP(07)05-04

J. Svensson, A Werner and JET EFDA contributors

# Large Scale Bayesian Data Analysis for Nuclear Fusion Experiments

# Large Scale Bayesian Data Analysis for Nuclear Fusion Experiments

J. Svensson, A Werner and JET EFDA contributors*

***JET-EFDA, Culham Science Centre, OX14 3DB, Abingdon, UK***

*Max Planck Institute for Plasma Physics, Teilinstitut Greifswald, Germany*
** See annex of M.L. Watkins et al, "Overview of JET Results ", (Proc. 21 st IAEA Fusion Energy Conference, Chengdu, China (2006)).*

Preprint of Paper to be submitted for publication in Proceedings of the IEEE International Symposium on Intelligent Signal Processing, Alcala de Henares, Madrid, Spain
( 3rd October 2007 - 5th October 2007)

**ABSTRACT.**
There is currently a paradigm shift taking place in the field of scientific methodology. Methods for the extraction of underlying physics from observations and the falsification/confirmation of scientific hypothesis are undergoing a significant change through the use of a generic approach to inference from observations: so called 'Bayesian' Probability Theory. The first part of this paper will outline and exemplify how this method is changing data analysis in nuclear fusion: How all uncertainties (systematic, statistical and model uncertainties) can be treated in a unified way, and how data analysis methods can be understood and unified through probability theory. The practical advantage here for nuclear fusion experiments is the possibility to utilise this method for a more comprehensive understanding of the internal state of fusion plasmas as inferred from measurements from multiple heterogeneous diagnostics. The second part of the paper will discuss architectural issues relating to the very complex analysis systems that might emerge from a systematic application of this method in large scientific experiments.

## 1. INTRODUCTION

Large nuclear fusion experiments pose a formidable data analysis challenge, related to the fact that a hot fusion plasma can seldom be directly probed by diagnostic equipment. The inference on internal parameters, such as particle densities, temperatures and electromagnetic fields, must therefore usually be made from quantities available only outside the plasma: radiation, escaped particles and external magnetic fields. This leads to comparably complex inversions, relying on detailed physics models for the processes that generate the measured quantities. It will be shown how by reformulating the inference problem of each diagnostic as a Bayesian inference on a common unknown physics 'state', measurements from different diagnostics can be combined for optimal inference on this physics state, without information losses. At the core is a new inversion method, Bayesian current tomography, to infer the internal magnetic field that provides the common coordinate system for the mapping of other diagnostic information. This model can then be fully integrated with individual Bayesian models of non-magnetic diagnostics to make optimal inference on further physics quantities.

The combination of a large number of diagnostic instruments for a joint Bayesian inversion leads to challenging architectural problems. A large nuclear fusion experiment has on the order of 100 diagnostic instruments and at least as many computer codes for modelling different aspects of the plasma. Each diagnostic is on its own a fairly complex piece of equipment with associated statistical and systematic uncertainties, viewing geometry, calibration procedures and associated physics models. The second part of this paper deals with the architectural problems that arise from a consistent application of this method of joint Bayesian analysis in fusion experiments. There will be two parts to the architecture discussed: The first part is a framework that allows a unified way of specifying general diagnostic models and their relationship to physics models. The second part is a framework dealing with unification of the calling, combination and parallelisation of modelling codes.

## 3. BAYESIAN INFERENCE

The foundation of so called 'Bayesian' probability theory is the association of an uncertain 'state of knowledge' with a probability or probability distribution. This view on probability [1,2] differs from the traditional 'frequentist' interpretation where a probability is viewed as the relative frequency of an event in the limit of an infinite number of trials. The Bayesian view accommodates a natural way of representing uncertainties that do not fit easily within a frequency interpretation, such as systematic uncertainties and model uncertainties. The real power of the Bayesian approach though, is that it seems to capture the very process of inference from data: the updating of one's knowledge about a given model or sets of models in the light of new observations, thereby making it a candidate for a theory of scientific inference in general [3].

The starting point of Bayesian inference on some unknown parameters of a model is a probability distribution $p(W)$, capturing the a *priori* knowledge one might posses about the otherwise unknown parameters, $W$, of a physics model. The information provided on the unknown parameters from some later observed data D, is then incorporated into the analysis through a second probability distribution that represents the likelihood of this particular data given any value of $W$: $p(D|W)$. These two pieces of knowledge, expressed as probability distributions, can now be combined using standard probability theory, to give the *posterior* distribution of W, representing what can be known about the model parameters after the observation of this particular dataset:

$$p(W|D) = \frac{p(D|W)p(W)}{p(D)} \tag{1}$$

where $(D) = \int p(D|W)\, p(W)\, dW$ is a normalisation constant independent of W. Equation (1) now captures everything we can know about the model parameters after the data has been observed, and is referred to as Bayes formula.

We now assume that we have a large number of diagnostics, collecting different type of data simultaneously during an experiment. Each diagnostic provides a set of measurements $D_i$ that provides information on different, possibly overlapping, parts of the vector of unknown physics parameters $W$ – the unknown physics–'state' of the plasma. We will now have one single prior distribution, $p(W)$, but several likelihood distributions, one for each $D_i$ : $p(D_i|W)$. If we now assume that the measurement noise on all measurements from all diagnostics are independent (that is, given a particular physics state $W$, the measurements are independent), the total likelihood will become

$$p(\{D_i\}|W) = \prod_{i=1}^{N} p(D_i|W) \tag{2}$$

and the posterior distribution on the set of all unknowns will be:

$$p(W \setminus \{D_i\}) = \frac{p(W)\prod_{i=1}^{N} p(D_i \setminus W)}{p(\{D_i\})} \tag{3}$$

which is the formula for integrated analysis of multiple diagnostics. This posterior distribution captures all that can be known about the physics parameters given the information from all diagnostics taken together, capturing interdependencies that occur because each diagnostic might give simultaneous evidence on overlapping parts of the vector W. Two examples of Bayesian diagnostic modeling for fusion diagnostics can be found in [4,5], and implementations of integrated Bayesian modeling in [6,7,8].

In practice, the vector $W$ might include so called nuisance parameters, uncertain parameters (such as calibration factors, geometrical parameters etc – usually regarded as–systematic uncertainties) that contribute to the uncertainty of our model parameters, but otherwise are not interesting to us. If we single those out by setting $W = \{W_M, W_N\}$ ($W_M$ 'model' parameters,' $W_N$ - 'nuisance' parameters), we can calculate their influence on our model parameters by integrating the posterior over the nuisance parameters:

$$p(W_M | \{D_i\}) = \int\int p(W_M, W_N | \{D_i\}) dW_N \tag{4}$$

Usually, such marginalisations over parameters will have to be carried out also over part of the model parameters, to arrive at 1D or 2D marginal distributions of quantities of interest. Such marginal distributions will have captured everything that can be known about those parameters, taking into account the uncertainties of all other free parameters in the system, whether they represent nuisance parameters or physics parameters. Such high dimensional integrals over parts of the posterior distribution usually represents the main computational effort of a Bayesian scheme, and stochastic integration techniques such as Gibbs sampling or Markov Chain Monte Carlo (MCMC) often have to be employed. Often though, analytical approximations of the posterior distribution around the maximum posterior value (MAP) can give satisfactory results.

## 3. BAYESIAN ANALYSIS FOR NUCLEAR FUSION

A major stumbling block in the implementation of (3) for nuclear fusion experiments is that physics parameters usually need to be mapped to a common 'magnetic' coordinate system, defined by the magnetic field line geometry inside the plasma. Since the internal magnetic field is continuously changing, it itself has to be inferred from (external) measurements. Through this mapping, the inferred coordinate system thus links a large number of diagnostics, and those other diagnostics therefore indirectly provide information on the magnetic geometry, even if their traditional 'role' is to measure some other physics parameter. To provide a reconstruction of the magnetic geometry that will also accommodate such interdependencies we have developed a

## 4. ARCHITECTURE FOR INTEGRATED ANALYSIS

Even though the Bayesian scheme for integration of diagnostic measurements (3) looks clean and simple, the reality behind the likelihood functions in (2) can be tremendously complex. It usually

includes the *forward function* of the diagnostic, which maps the physics parameters to an expected measurement. This function includes all calculations connecting the physics of the plasma to the final measurement. It also includes all types of uncertainties for that specific diagnostic: calibration factors, positioning parameters etc. The more complex diagnostics can have dependencies on hundreds of external parameters.Assumptions at any level of these models, from physics to the positioning of diagnostic lines of sight, can influence the final result in the joint analysis. For such a complex utilisation of available information to work and be manageable, a generic architecture is mandatory. The rest of this paper is therefore devoted to an overview of two architectural parts that have been developed to handle this complexity. The first, MINERVA, is a generic framework that formalises the specification of diagnostic models, parametric dependencies, and their relationship to physics models. This framework allows optimal seamless combination of diagnostics and optimal Bayesian inference on physics models evidenced by heterogeneous diagnostics incorporated into the framework.

All parametric dependencies of a diagnostic or a physics model are 'visible' to the framework and allows a decoupling between diagnostic models, physics models, data sources, and specific inversion algorithms (such as MAP optimization, MCMC etc). This makes it possible to have control over every single parameter in the combined model, and through the usage of well defined interfaces, underlying physics models can even be swapped. Figure 3 shows the main principle behind the workings of this framework, which has been used to implement the current tomography described in section 3. Forward Mode Inverse Mode.

The unified handling of diagnostic models as described above is a solution to only one part of the problem. The other has to do with the many different physics calculations that need to be performed for the definition of the physics models in the framework. These are often already implemented in large legacy codes, typically written as stand-alone FORTRAN programs. For example, the current tomography model relies on a 3D magnetostatic code (MAG3D), written by the authors, to calculate magnetic field and vector potentials from different types of current elements. We believe that unification of the process of the programmatic interaction with these codes would lead to advantages also outside the scope of Bayesian analysis, since those codes would be easily accessible from any environment for other uses. The practical process of accessing, understanding and running these legacy codes is currently complicated and creates a barrier for the general user, not an expert of a specific code. The possibility of being able to combine such codes with each other, the output of one code being used as input or partial input to other codes – possibly in an iterative fashion, would lead to much faster development and scripting of advanced modelling scenarios. Currently, to combine just a couple of codes with each other could amount to months or even years of work. The SOFI (Service Oriented Fusion Initiative) project is trying to remedy this by building codes into a so called 'Service Oriented' Architecture (SOA), using Web Service technologies. In this model, all data communicated between codes are strictly typed in language independent XML Schema (XSD) structures and the functional interface between codes and between

users and codes are defined using language independent XML WSDL (Web Service Description Language) documents [10]. In this way the complete interaction with a code can be defined in a language-independent and machine-readable way, so that caller stubs for a large number of languages (possibly different from the one in which the legacy code was originally written) can be automatically generated from such definitions.

This creates virtually endless possibilities for instantaneous utilisation and combination of legacy codes (callable from C/C++, Java, Python, Matlab, Mathematica etc) for complex high-level calculations. Also, parallelisation of these codes in a simple fashion becomes possible. The handling of code and data resources in SOFI is done through an internally developed architecture (a 'micro-grid') for localisation, management and parallelisation/load distribution of codes exposed as web services. Any code encapsulated as a web service, with its interface defined through a WSDL document, can be handled by the framework, whose internal components are also themselves implemented as web services. Figure 4 shows the usage of the framework through a WebService Broker that deals with requests from clients for single or multiple instances of given services/ applications. This broker is itself a web service and can therefore be reached from virtually any programming language or environment. It will handle load balancing issues through communication with optional LoadProbes at each server computer. Figure 5 shows the part of the SOFI framework dealing with optional managed' web services, whose life cycle can be controlled by the NetDisponent service, which is also a web service. Currently codes for magnetostatic calculations, field line tracing, finite element representation of CAD models, generic XML machine component database, and interface to equilibrium codes have been incorporated into the SOFI architecture. It has been tested for parallel calculations on up to 112 distributed processors.

## CONCLUSIONS

The increase in complexity of both diagnostic equipment and physics models in large scientific experiments poses a problem at least as large as the more commonly observed problem of handling and storage of data, and is directly related to the possibility of testing scientific models in such experiments. We have presented a principled approach to large scale analysis using Bayesian inference as the overarching method, applied it to inference on the central magnetic coordinate system of fusion plasmas, and outlined the two architectural principles we are applying for the handling and implementation of the complex relationship between models and diagnostic measurements that exist in modern nuclear fusion experiments.

## REFERENCES

[1]. D. Sivia, J. Skilling, Data Analysis: A Bayesian Tutorial, Oxford University Press, 2006

[2]. A. Gelman, J.B Carlin, H.S Stern, D.B Rubin, Bayesian Data Analysis, Chapman & Hall, 2003

[3]. C. Howson, P Urbach, Scientific Reasoning: The Bayesian Approach, Open Court Publishing Co, 1993

[4]. R. Fischer, A. Dinklage, E. Pasch, “Thomson scattering analysis with the Bayesian probability theory”, Plasma Phys Controlled Fusion, Vol **44** No 8, 2002

[5]. J. Svensson, R.W.T König, ”Bayesian Modelling of Spectrometer Systems”, 32nd EPS Conference on Plasma Physics, 2005

[6]. J. Svensson, A. Dinklage, R. Fischer, “An integrated Data Analysis Model for the W7-AS Stellarator”, 30th EPS Conference on Plasma Physics, 2003

[7]. J. Svensson, A. Dinklage, R. Fischer, J. Geiger, A. Werner, “Integrating Diagnostic Data Analysis for W7-AS using Bayesian Graphical Models”, Rev Sci Instrum Vol **75**, p. 4219-4221, 2004

[8]. A. Dinklage, C.D. Beidler, R. Fischer, H. Maasberg, J. Svensson, Y. Turkin, “Integrated Interprative Transport Modelling”, 31th EPS Conference on Plasma Phyicss, 2004

[9]. L.L. Lao, H St John, R D Stambaugh, “Reconstruction of current profile parameters and plasma shapes in tokamaks”, Nuclear Fusion **25** (1985), pp. 1611-1622

[10]. Web Service Description Language (WSDL) Version 2.0, http://www.w3.org/TR/wsdl20/

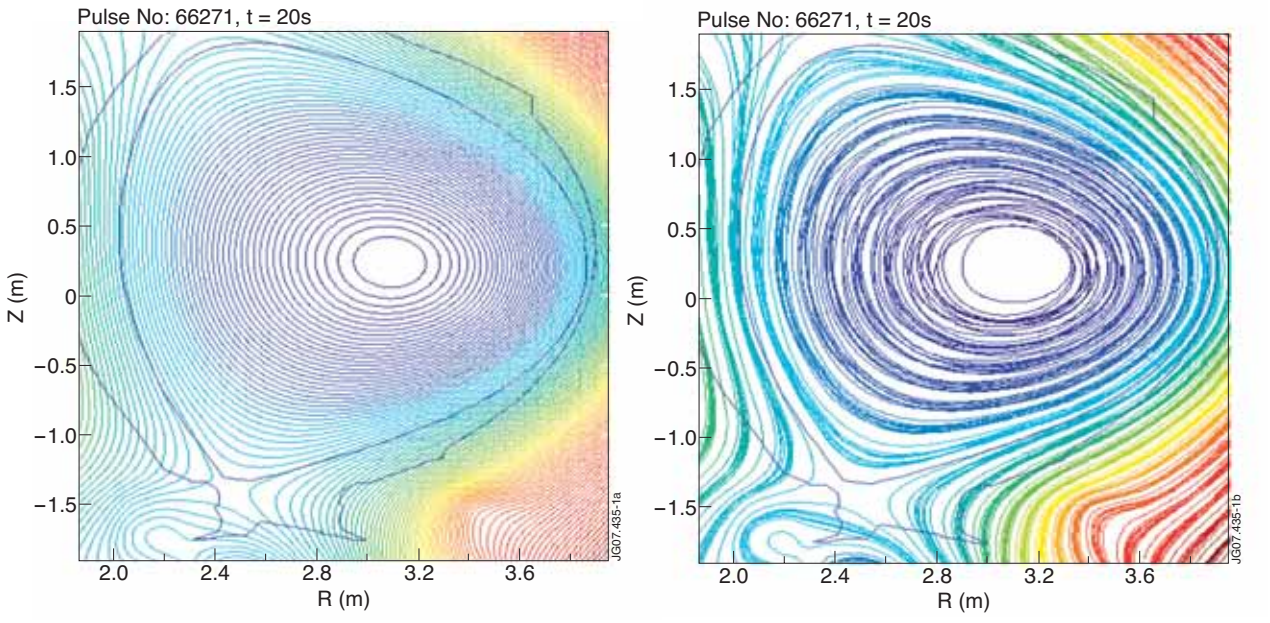


*Figure 1: a) Maximum posterior (MAP) estimate of surfaces of constant magnetic flux, b) samples from the posterior distribution showing the uncertainty of the reconstruction. The widths of the contour lines correspond to the uncertainty at that position of the common magnetic coordinate system. Plasma boundary line is taken from another method (force balance – EFIT [9]) for comparison.*

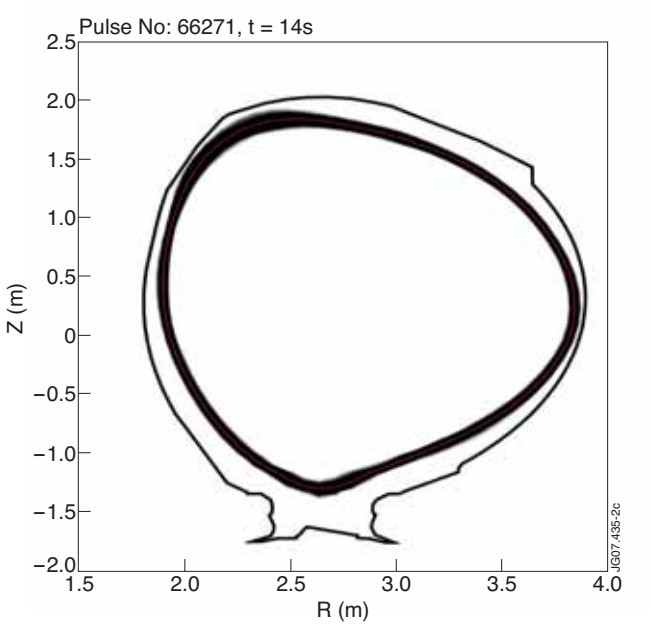


*Figure 2: The plasma boundary and uncertainties as reconstructed from the current tomography method. The higher uncertainty in the upper left part is likely attributable to the proximity of the plasma to the magnetized transformer iron core, whose unknown magnetization contributes to the uncertainty of the reconstructed boundary.*

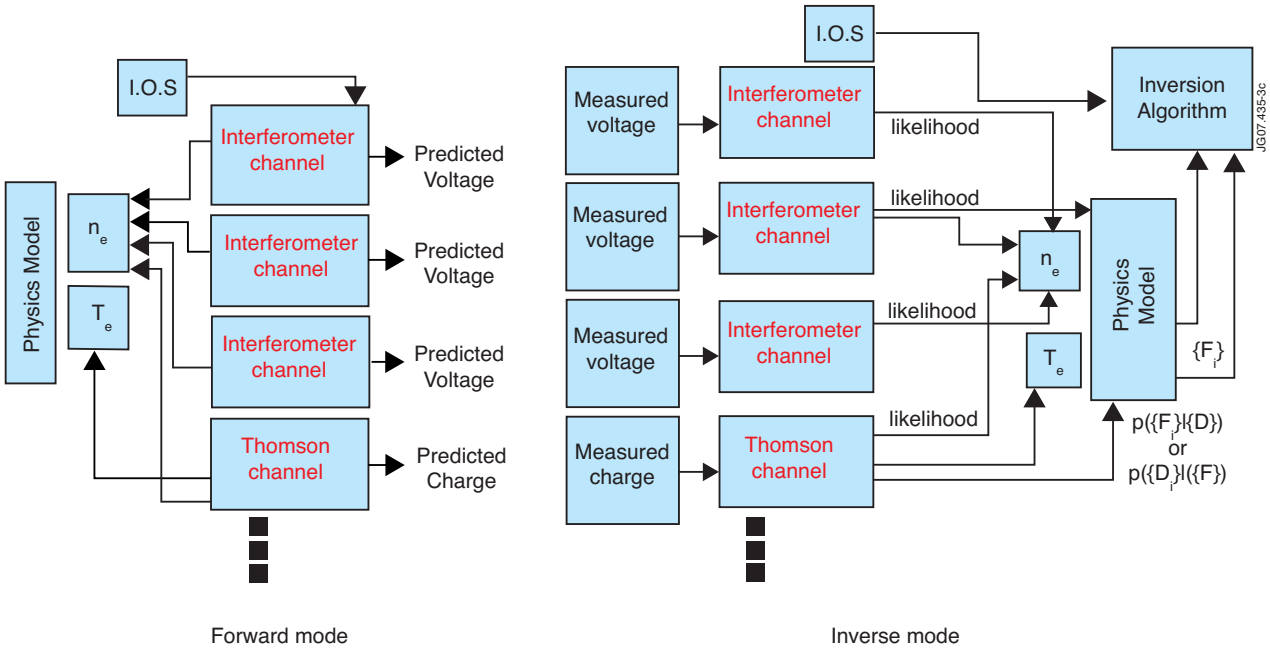


*Figure 3: The principle behind the MINERVA framework for Bayesian analysis. Diagnostics are decoupled from physics models and can be arbitrarily combined. In the 'forward mode', expected measurements of participating diagnostic can be calculated from underlying joint physics models. In the 'inverse mode', physics parameters are optimally inferred from the combined set of connected diagnostics.*

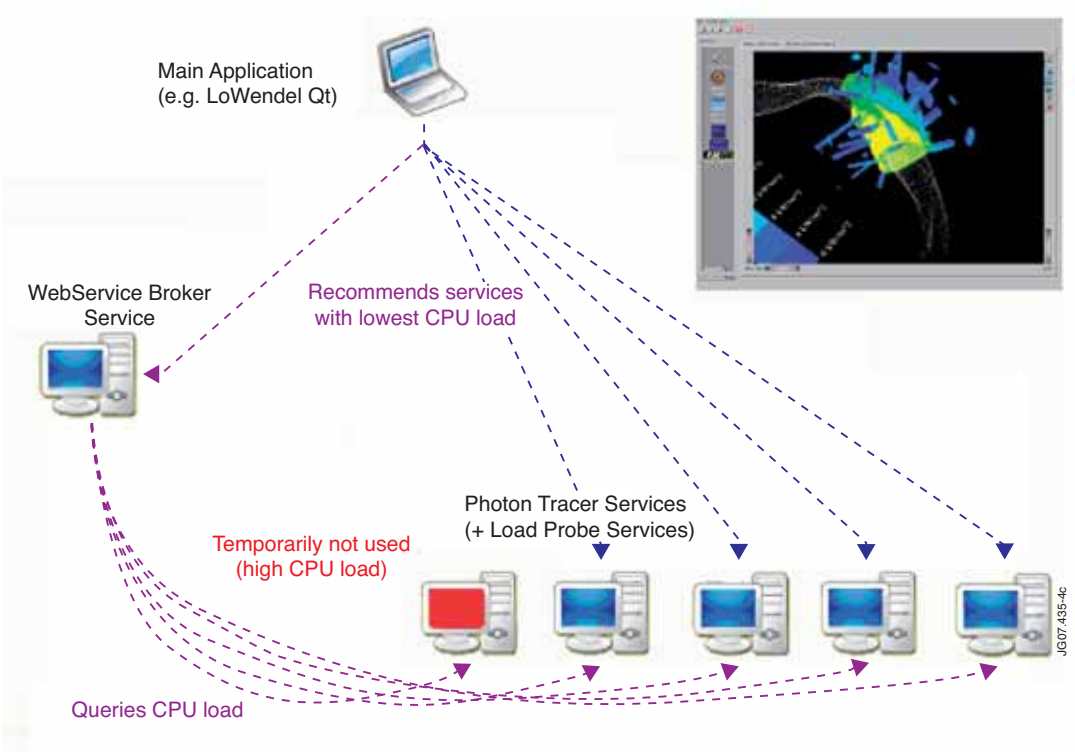


*Figure 4: Usage of the SOFI framework through a web service broker , that manages access, load balancing and parallellisation of participating web service applications.*

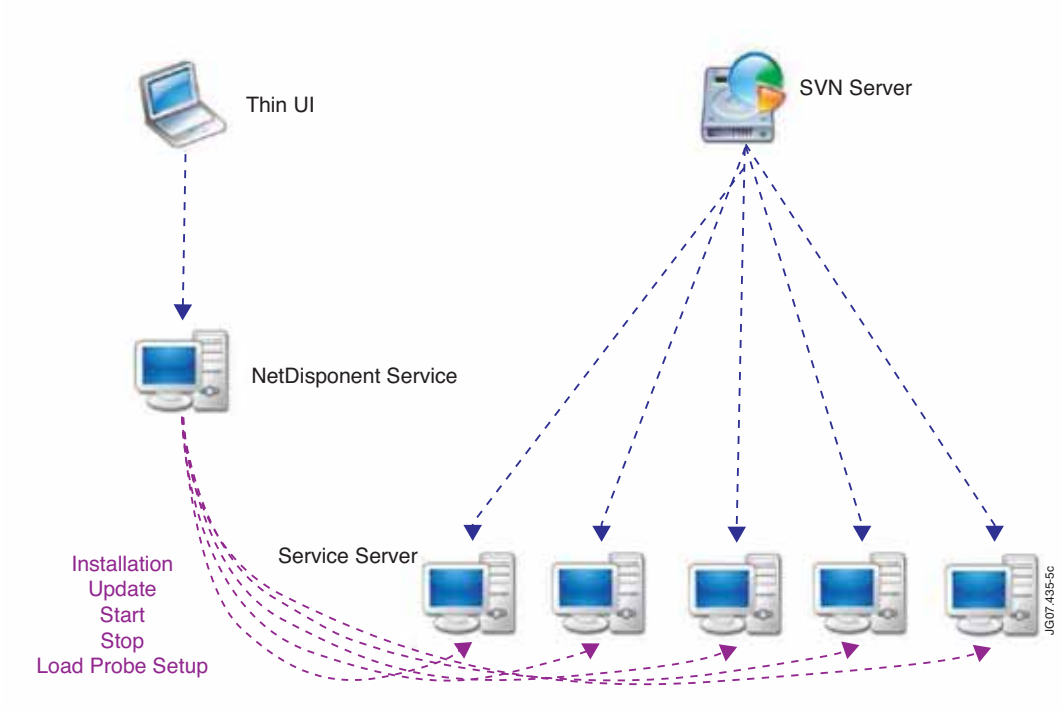


*Figure 5: Part of the SOFI framework for managing the lifecycle of services through the NetDisponent web service.*