

HPC for Bioinformatics



Blueprint



Center of
Excellence



A program of the Samuel Lunenfeld
Research Institute

A University of Toronto affiliated
research institute



Table of Contents

- ◆ What is Blueprint?
- ◆ Sun Technologies Inside Blueprint
- ◆ JAXB – accelerated Java development
- ◆ HPC Web Services
- ◆ HPC Data Processing
- ◆ What's next?

What is Blueprint?

- ◆ Led by principal investigator Dr. Christopher Hogue
- ◆ **Bioinformatics** research program in the Samuel Lunenfeld Research Institute inside Mount Sinai Hospital
- ◆ Affiliated with the University of Toronto
- ◆ SUN Microsystems Center of Excellence
- ◆ Secured government and public funding of \$29 million over 3 years
- ◆ 68 employees in Toronto, 6 in Singapore
- ◆ Operational for ~2 years

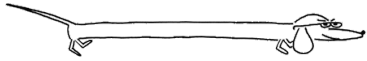
What is Bioinformatics?

- ◆ The application of computer technology to the management of biological information
- ◆ Software applications used to gather, store, analyze and integrate biological information
- ◆ Databases and algorithms designed for the purpose of enhancing the process of biological research

Sun Technologies inside Blueprint

- ◆ 6 million (USD) worth of Sun hardware and software
 - ◆ 108 node HPC cluster
 - ◆ 16 Sun Servers (280,480,880,1280)
 - ◆ 24 TB (2x 3960) of storage
 - ◆ iPlanet Tools (LDAP, Mail, Calendaring)
- ◆ JAXB + other development tools

Software Development



seqhound.blueprint.org

◆ SeqHound

- ◆ Database containing sequences and structures
- ◆ Freely available local & remote APIs available in several languages



bind.ca

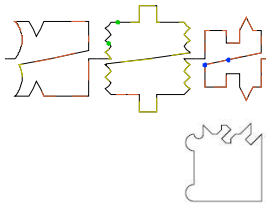
◆ BIND

- ◆ Database of interaction, molecular complex and pathway information
- ◆ Freely available web application designed to create, edit and retrieve records



◆ MMDBBIND

- ◆ Subset of BIND containing structurally based interactions
- ◆ Interactions automatically determined from PDB records



◆ Proteoglyphs / Ontoglyphs

- ◆ Proteoglyphs are symbols used to represent protein domains
- ◆ Ontoglyphs are symbols that encode function and location of a protein

◆ Distributed Folding

- ◆ Distributed computing approach to the study of protein folding



www.distributedfolding.org

BIND Software Past & Present

	“Classical” BIND	Current BIND
Web Interface	C CGI / PHP	JSP / Servlets
Middleware	C Libraries	Java Packages
Database Layer	Codebase/ DB2 CLI	JDBC/SQL
Data Specification	ASN.1	XML Schema
Code Generator	Asntool/Datatool	JAXB
Text Indexing	In-house Text Indexer	Lucene Derivative
Architecture	“Classic C Style”	

BIND Specification

- ◆ Describes interactions, complexes and pathways
- ◆ Describes proteins, DNA, RNA, genes, small molecules
- ◆ Over 2000 data fields
- ◆ Built on top of the NCBI specification
- ◆ Available in ASN.1 and XML Schema
- ◆ Published in NAR (2003)
- ◆ Currently at version 3.1

JAXB Accelerated Development

Going From 2000 Data Fields to 750K Lines of Code
SUN's Java Architecture for XML Binding (JAXB)

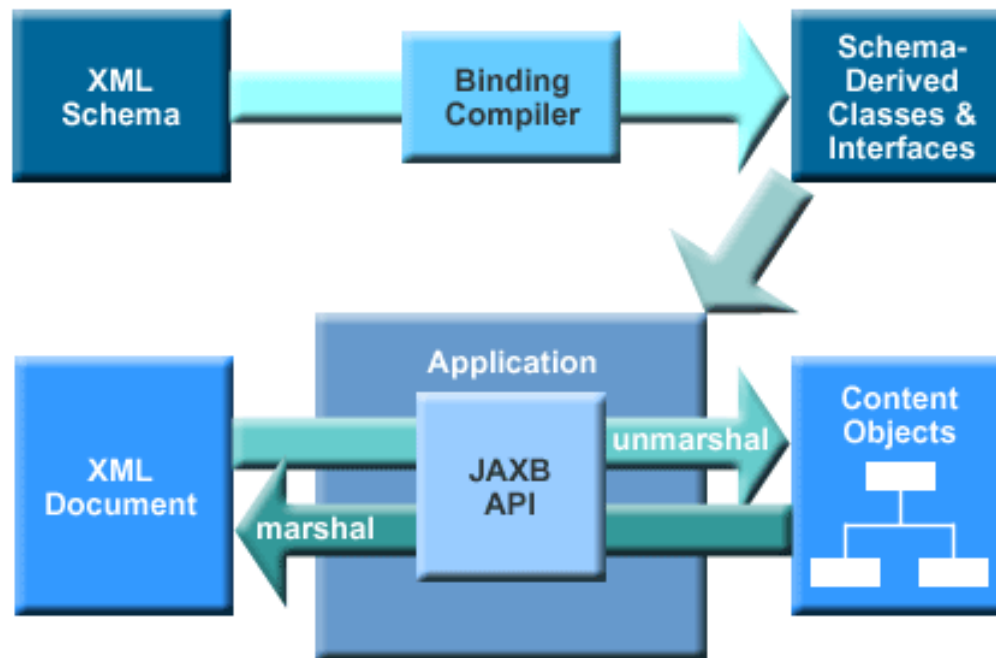
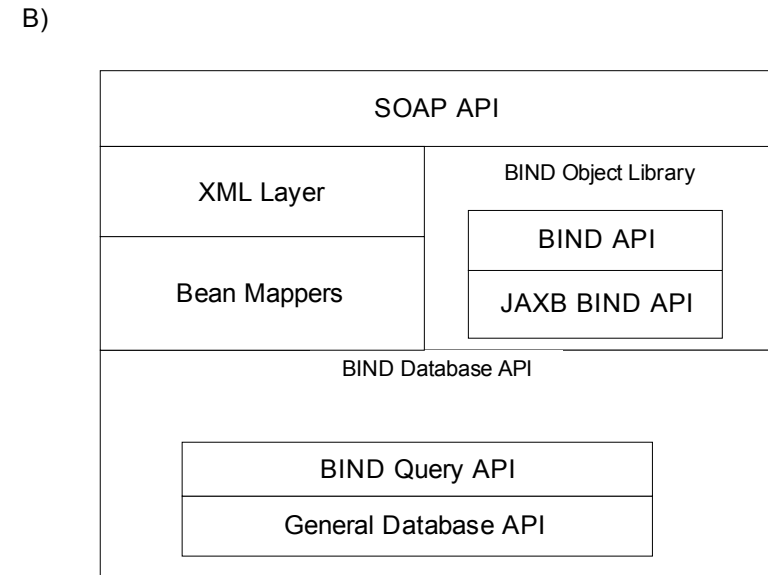
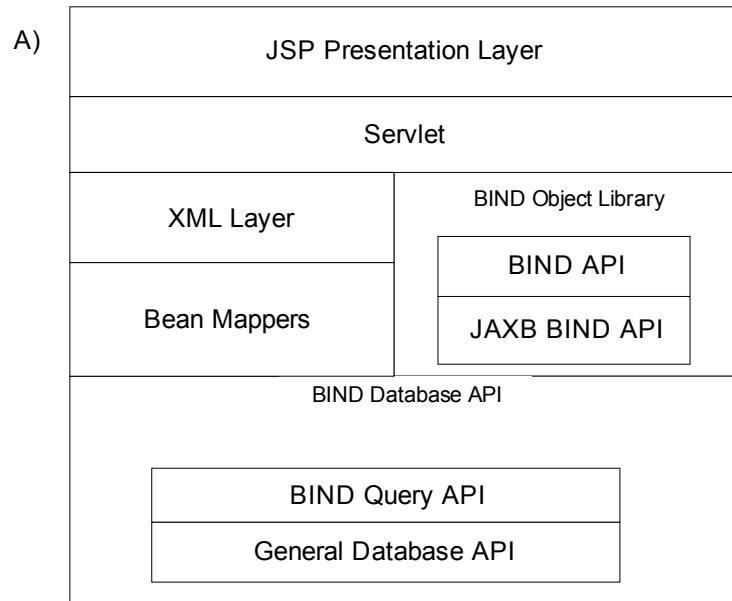


Image taken from <http://java.sun.com/developer/technicalArticles/WebServices/jaxb/index.html>

Why J2EE & XML ?

- ◆ Well defined web application development practices
- ◆ Allows use of SQL compliant databases interchangeably
- ◆ Rapid web application development
- ◆ Increased ease of maintenance
- ◆ Inherent security policies
- ◆ Widely used (i.e. many 3rd party developers / documentation / resources)
- ◆ Backed by SUN, IBM & W3C

BIND Software Architecture



HPC Clustering

- ◆ 108 Dual 2.8ghz V60x nodes
- ◆ 4GB RAM per node
- ◆ Gigabit over Fibre to Foundry BigIron switches
- ◆ RedHat 8.0 customized for bioinformatics applications
- ◆ 2 TB 'scratch' space for developers to run misc. applications



HPC for Data Processing

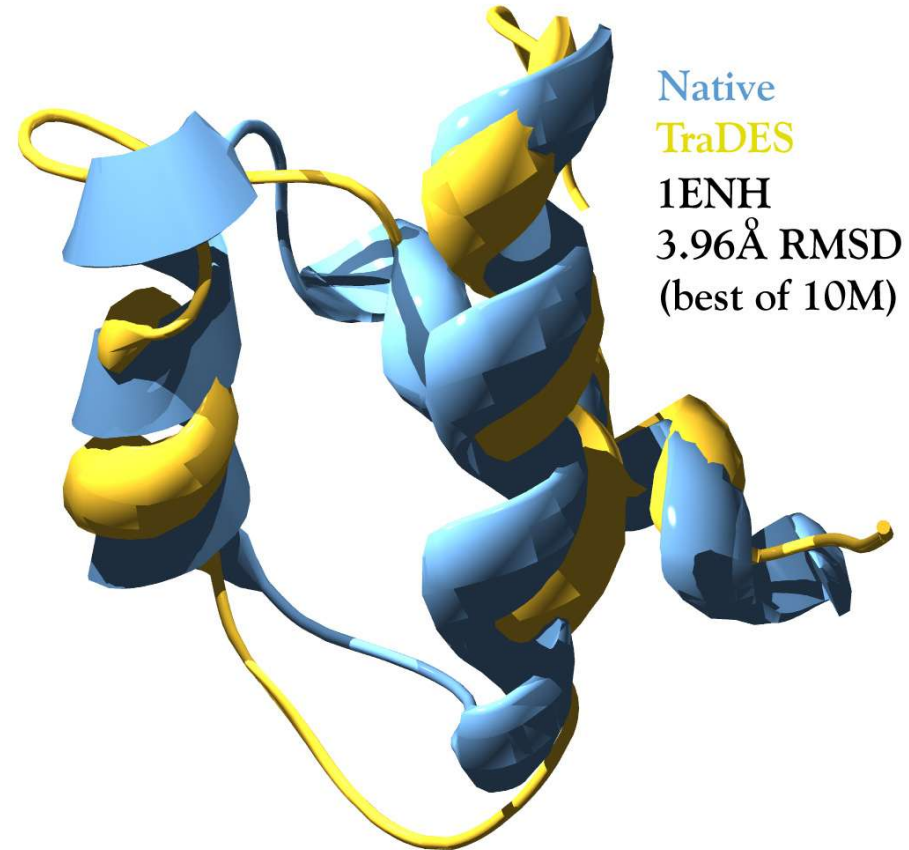
- ◆ nblast and rpsdb
 - ◆ Customized versions of Blast and RPS to run on the HPC cluster at Blueprint
- ◆ nblast does an NxN (omitting dupes) comparison against the entire contents of our database
- ◆ rpsdb does NxM comparison of all sequences in a sequence database against all conserved domains in a CD database
- ◆ Both store results into another database
- ◆ nblast takes 6 days – 310 million records produced
- ◆ rpsdb takes 3 days – 29 million records produced
- ◆ Results of these jobs are available via FTP from our FTP site

Distributed HPC

- ◆ <http://www.distributedfolding.org/>
- ◆ An attempt to harness the power of the Internet to perform massive brute force sampling of protein structures
- ◆ Could get as many as 10 billion samples of a 100-residue protein in one month, expected to produce at least one structure within 6-8Å of the native fold

Distributed Folding Stats

- ◆ Over 3,000 active users
- ◆ Over 10,000 nodes actively processing data
- ◆ Approx 5 teraflops CPU processing power available



What Next?

- ◆ New office in Singapore
 - ◆ Will clone Toronto facility
 - ◆ Global load balancing between sites
- ◆ New office in Europe
- ◆ SOAP API's for BIND, and soon Seqhound
- ◆ OBDC version of Seqhound
- ◆ New visualization software