Introduction to Volunteer computing and BOINC

Nicolas Maire, Swiss Tropical Institute

with François Grey and Ben Segal, CERN

Objectives and Goals

- Overview of volunteer computing and BOINC
- BOINC concepts and components
- Hands-on:
  - BOINC client and web-interface
  - BOINC server

Schedule Thursday, November 15th

- 9-10.30 Overview of Volunteer computing and BOINC
- 11-12.30 Hands-on 1:
  -Client
  -Server prerequisites
- 14-15.30 BOINC Advanced: Client and Server
- 16-17.30 Hands-on 2:
  -BOINC Server setup

Overview BOINC I

- Volunteer computing
- Volunteer vs. Grid Computing
- BOINC
  - Architecture
  - Communities
  - malariacontrol.net: A BOINC project example
  - -Outlook

What is Volunteer computing?

- Volunteer computing (VC) is arrangement in which people (**volunteers**) provide computing resources to **projects**, which use the resources to do distributed computing.
- Projects are typically academic (university-based) and do scientific research
- Volunteers are typically members of the general public who own Internet-connected PCs.
- Millions of people are donating spare time on their computers for scientific projects. Anyone with a computer and Internet-access can join.
- The first volunteer computing project was GIMPS (Great Internet Mersenne Prime Search), which started in 1995. Other early projects include distributed.net, SETI@home, and Folding@home.
- Today there are dozens of active projects.

Is VC a form of Grid computing?

- Both are forms of distributed computing that try to more fully utilize existing resources.
- Both enable distributed computing on a global scale
- Both are adapted to massively parallel computing

- However, they differ in several essential respects

Is VC a form of Grid computing?

- Unlike Grids, there is no mutual accountability between partners in Volunteer computing
- Volunteers are effectively anonymous
  - Software for volunteer computing must accommodate the possibility of misbehavior
- Volunteers must trust projects in several ways
  - Applications that don't damage their computer or invade their privacy
  - The project is truthful about what work is being done, and how the results will be used
  - The project follows proper security practices, so that hackers cannot use the project as a vehicle for malicious activities.

Is VC a form of Grid computing?

- Volunteer computing “pulls”; it does not “push”
  - Requires the use of a “pull” model in which PCs periodically request work from a central server, rather then the “push” model used by most grid software.
- Volunteer computing uses the “commodity Internet”
  - Both projects and volunteers must pay for network bandwidth. Data-intensive applications require careful planning.
- Volunteer computing must embrace amateurs
  - Volunteered resources are owned by regular people, not by IT professionals. The software must be simple to install and run.
- Volunteer computing demands great public relations
  - Scientists can access volunteer computing power not by requesting or purchasing allocations, but by persuading the public that their research is worthwhile. Public outreach is a significant fringe benefit of Volunteer computing.

Desktop grid computing

- A form of distributed computing in which an organization uses its existing desktop PCs to handle its own long-running computational tasks
- Superficially similar to volunteer computing, but because it has accountability, it is significantly different
- The computing resources can be trusted. No need for redundant computing
- Client deployment is typically automated
- Although originally designed for volunteer computing, BOINC works well for desktop grid computing

Volunteer computing projects by field

SCIENCE

SETI@home (**BOINC**)
evolution@home
eOn
climateprediction.net (**BOINC**)
Muon1
LHC@home (**BOINC**)
Einstein@Home(**BOINC**)
BBC Climate Change Experiment (**BOINC**)
Leiden Classical (**BOINC**)
QMC@home (**BOINC**)
NanoHive@Home (**BOINC**)
µFluids@Home (**BOINC**)
Spinhenge@home (**BOINC**)
Cosmology@Home (**BOINC**)
PS3GRID (**BOINC**)
Mars Clickworkers

LIFE SCIENCES

Parabon Computation
Folding@home
FightAIDS@home
Übero
Drug Design Optimization Lab (D2OL)
The Virtual Laboratory Project
Community TSC
Predictor@home (**BOINC**)
XGrid@Stanford
Human Proteome Folding (**WCG**)
CHRONOS (**BOINC**)
Rosetta@home (**BOINC**)
RALPH@home (**BOINC**)
SIMAP (**BOINC**)
malariacontrol.net (**BOINC**)
Help Defeat Cancer (**WCG**)
TANPAKU (**BOINC**)
Genome Comparison (**WCG**)
Docking@Home (**BOINC**)
proteins@home (**BOINC**)
Help Cure Muscular Dystrophy (**WCG**)

MATHEMATICS AND CRYPTOGRAPHY

Great Internet Mersenne Prime Search
Proth Prime Search
ECMNET
Minimal Equal Sums of Like Powers
MM61 Project
3x + 1 Problem
Distributed Search for Fermat Number Divisors
PCP@Home
Generalized Fermat Prime Search
PSearch
Seventeen or Bust
Factorizations of Cyclotomic Numbers
Goldbach Conjecture Verification
The Riesel Problem
The 3*2^n-1 Search
NFSNET
Search for Multifactorial Primes
15k Prime Search
ElevenSmooth
Riesel Sieve
The Prime Sierpinski Project
P.I.E.S. - Prime Internet Eisenstein Search
Factors of k*2^n±1
XYYXF
12121 Search
2721 Search
Operation Billion Digits
SIGPS
Primesearch
Lone Mersenne Hunters
Factoring
100 Million digits prefactor project
Repdigit Prime Problems
Mersenneplustwo Factorizations
Sierpinski/Riesel Base 5
SZTAKI Desktop Grid (**BOINC**)
Riesel Prime Search
Proth Sieve
Twin Internet Prime Search
Pi Segment
Rectilinear CN (**BOINC**)
ABC@home (**BOINC**)
WEP-M+2 Project (**BOINC**)
distributed.net
PrimeGrid (**BOINC**)
M4
HashClash (B**OINC**)
Assault on 13th Labour
Free Rainbow Tables

Source: http://distributedcomputing.info/

Volunteer Computing

- Because of the huge number of PCs in the world, volunteer computing can (and does) supply more computing power to science than does any other type of computing.
- This advantage will increase over time, because consumer electronics (PCs and game consoles) will advance faster than more specialized products, and that there will simply be more of them.
- Volunteer computing encourages public interest in science, and provides the public with voice in determining the directions of scientific research.

Volunteer Computing Performance

- folding@home (non-BOINC) passed petaFLOPS mark in September 2007
- Using CPUs, GPUs
- Runs on and is distributed with Sony PS 3

- BOINC combined around 600 teraFLOPS

- IBM's Blue Gene/L at 360 teraFLOPS in September 2007

What is BOINC?

- Berkeley Open Infrastructure for Network Computing
- Software platform for distributed computing using volunteered computer resources
- http://boinc.berkeley.edu

BOINC features

- Project autonomy
  - Projects are independent; each one operates its own servers and databases. There is no central directory or approval process.
- Volunteer flexibility
  - Volunteers control which projects they participate in, and how their resources are divided among projects. When a project is down or has no work, the resources of its volunteers are divided among other projects.
- Flexible application framework
  - Existing applications in common languages (C, C++, Fortran) can run as BOINC applications with little or no modification. New versions of applications can be deployed without required any action by volunteers.

BOINC features

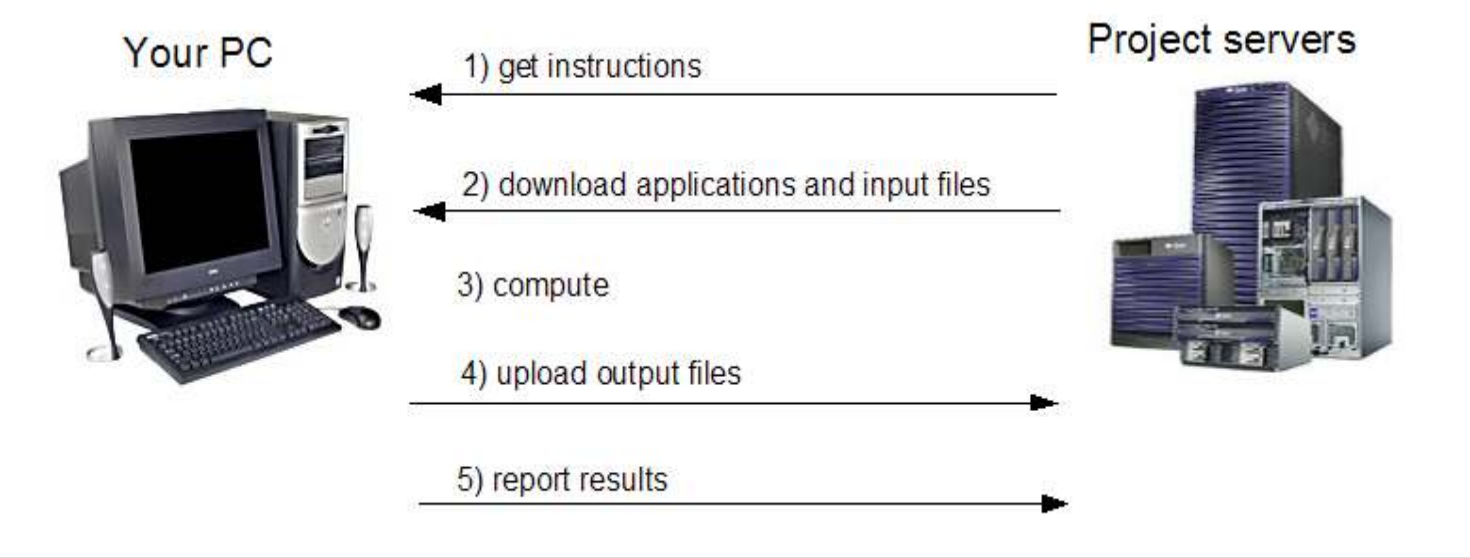
- Security
  - BOINC protects against several types of attacks. For example, digital signatures based on public-key encryption protect against the distribution of viruses.
- Multiple participant platforms
  - The BOINC core client is available for most common platforms.
- Open, extensible software architecture
  - BOINC provides documented interfaces to many of its key components, making it possible for third-party developers to create software and web sites that extend BOINC.
- Volunteer community features
  - BOINC provides web-based tools, such as message boards that encourage volunteers to form online communities.

BOINC features

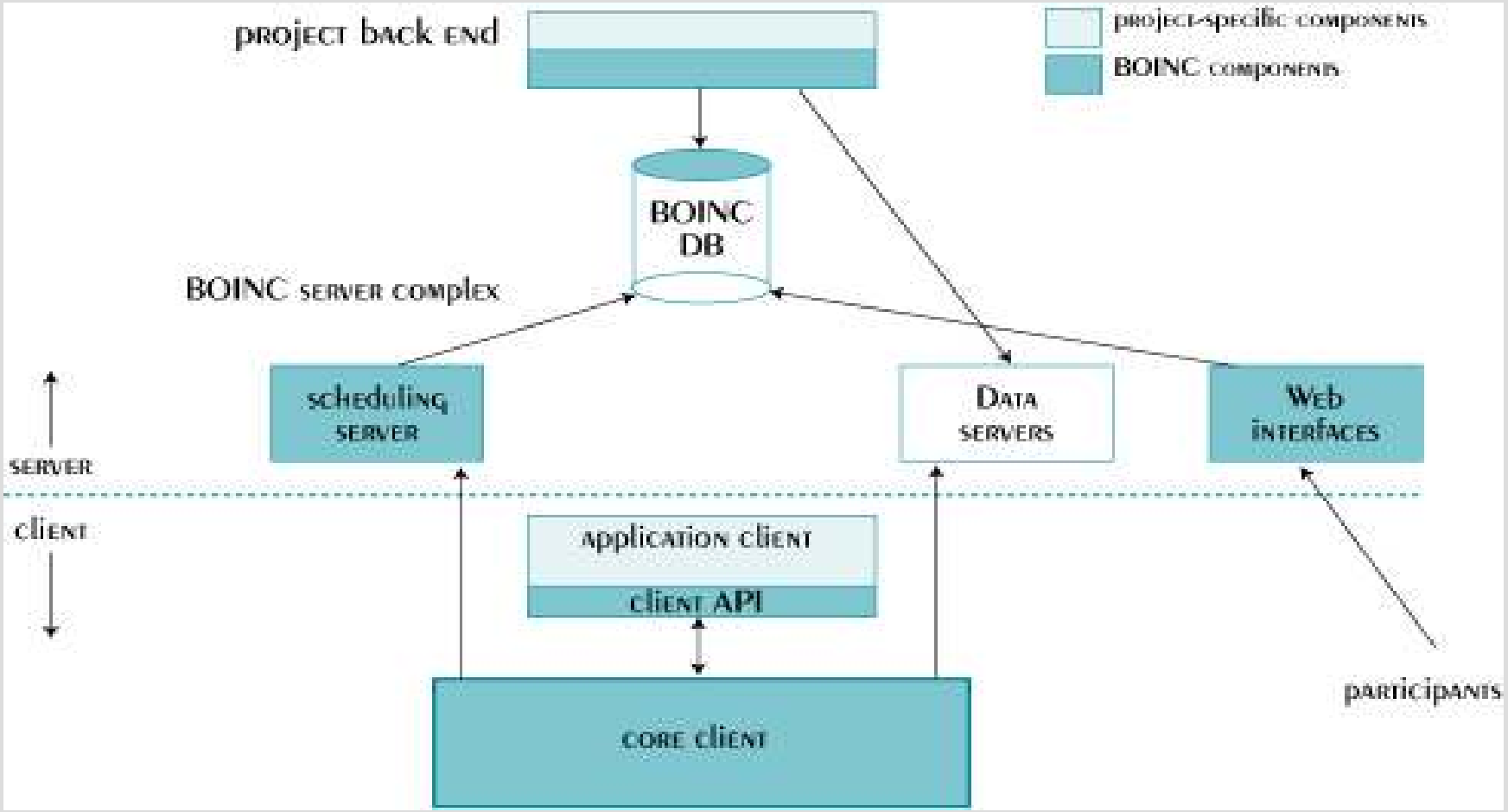
- Multiple participant platforms
  - The BOINC core client is available for most common platforms.
- Open, extensible software architecture
  - BOINC provides documented interfaces to many of its key components, making it possible for third-party developers to create software and web sites that extend BOINC.
- Volunteer community features
  - BOINC provides web-based tools, such as message boards that encourage volunteers to form online communities.

Basic structure of BOINC

- Interaction between **client and server**



Basic structure of BOINC



BOINC client

- Available for several computer platforms (Windows, Mac, Linux,…)
- User download from boinc.berkeley.edu
- Attaches to projects and assigns quotas
- Optionally defines personal preferences
- That’s it, the client handles the rest

Incentives for volunteers

- Philanthropy
- Curiosity
- Fun (play, competition)
- Community (Message boards, teams)
- Prestige (credit, recognition)

BOINC credits

- Credit points are awarded for successful results
- Credits are an important incentive
  - For competitive individual users or teams
  - But users who participate to help science often also like to have something measurable in return for their donation
- Credits provide a useful, though imperfect, performance measure for projects (or BOINC as a whole)

How computations are credited

- BOINC's unit of credit is the Cobblestone (after Jeff Cobb of SETI@home)
- A Cobblestone is 1/100 of a day of CPU time on a reference computer (a computer which produces certain benchmark results)
- Hosts claim some amount of credit for every result they report to a project server
- The project grants credit to the host if the result is "validated"
  - Either a fixed amount if workunit time is predictable
  - Else the average of claimed credit values for that workunit

Credit listings

- Statistics panel in BOINC client
- “Your Account” web page
- “Top participants” web page
- BOINC statistics sites

- Listings usually distinguish between total credit and Recent Average Credit (RAC)

Stats sites

- For example, Willy de Zutter's BOINCstats
- Projects regularly export XML-dumps of the database status
- Stats sites collect and process these dumps
- Participants and hosts linked across projects using unique ids in the XML
  - Based on email address of user
  - Based on hash of host properties

BOINC security features

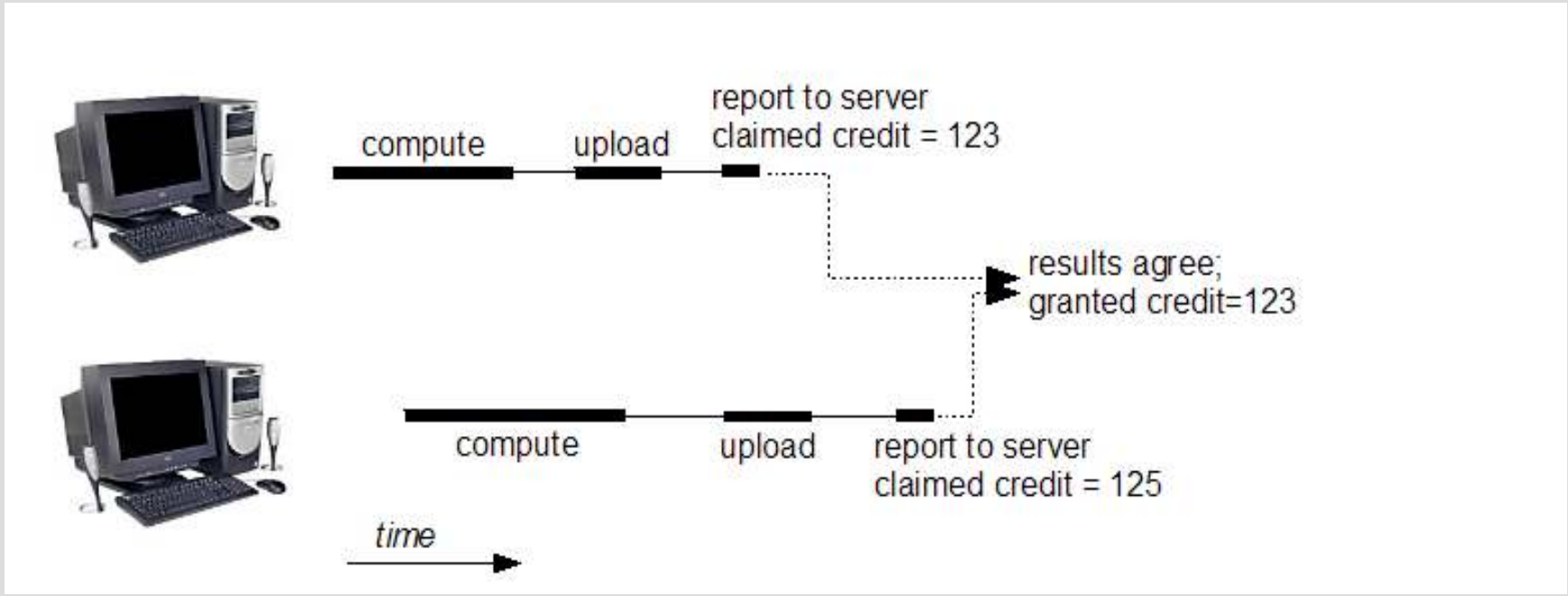
- BOINC uses code signing to prevent malicious executable distribution
- All files associated with the applications are sent with digital signatures

BOINC security features

- BOINC detects when applications use too much disk space, memory, or CPU time, and aborts them.
- BOINC also prevents denial of service attacks to the server, result falsifications and credit falsifications.
- But applications are currently only “sandboxed” on Mac OS X. Participants must understand that when they join a BOINC project, they are entrusting the security of their systems to that project.
- Sandboxing on Windows will come with client version 6

Redundant computing

- Identifying erroneous results and granting credit



Applications suitable for BOINC

- The main requirement of the application is that it be divisible into a large number (thousands or millions) of jobs that can be done independently.

- Additional requirements:
- **Public appeal**
  - An application must be viewed as interesting and worthwhile by the public. A project must have the resources and commitment to maintain this interest, typically by creating a compelling web site.
- **Low data/compute ratio**
  - Input and output data are sent through commercial Internet connections, which may be expensive and/or slow. If your application produces or consumes more than a gigabyte of data per day of CPU time, then it may be cheaper to use in-house cluster computing rather than volunteer computing.
- No dependence on short turnaround
  - There is no guarantee that results are returned within a certain time span

BOINC resource requirements

- BOINC-enabling an existing science application takes about three man-months: one month of an experienced sys admin, one month of a programmer, and one month of a web developer (rough estimates)
- Once the project is running, budget a 50% FTE (mostly system admin) to maintain it
- In terms of hardware, you'll need a mid-range server computer, the requirements are highly project-specific
- You'll also need a fast internet connection

A BOINC project example: malariacontrol.net

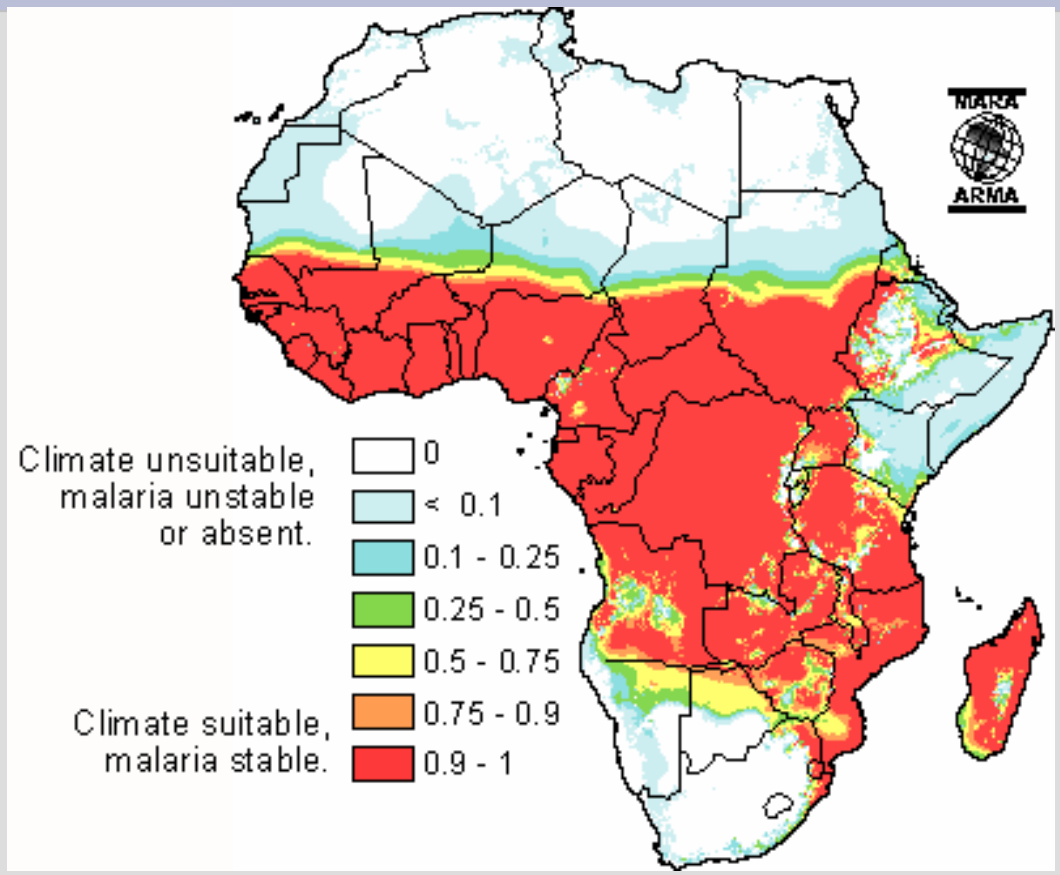
Simulation of malaria epidemiology

- **Simulation models** of transmission dynamics and health effects of malaria are an important tool for malaria control.
- Models help develop **optimal strategies** for delivering mosquito nets, chemotherapy, or new vaccines currently under development and testing.
- Such **modelling is computer intensive,** requiring simulations of large human populations with diverse parameters related to biological and social factors that influence disease distribution.

.

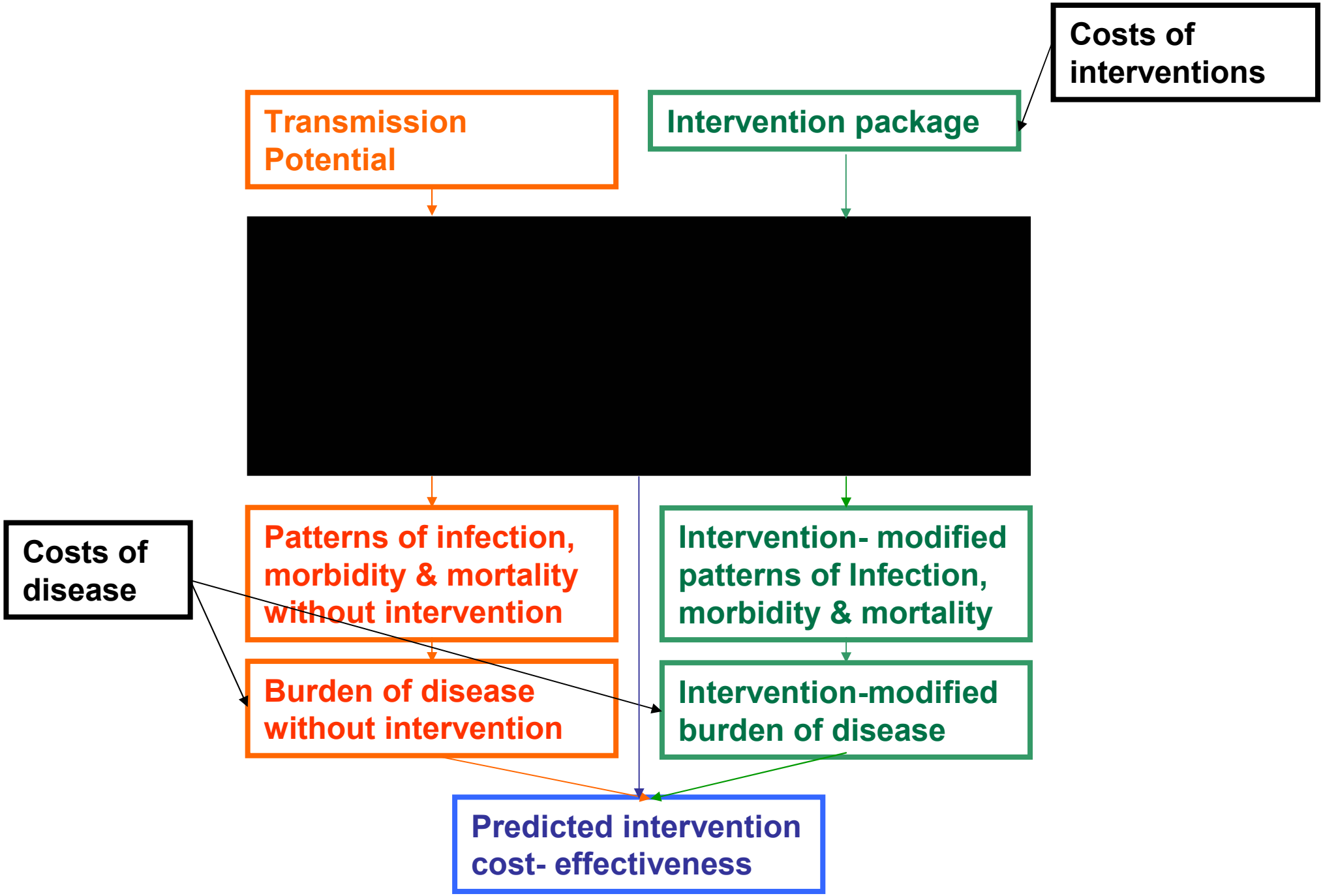
Why malaria models?

- Along with HIV/AIDS, malaria one of the two most important health problems in Africa
- Causes hundreds of millions of episodes of illness each year, and over 1 million deaths
- Up to 40% of health expenditure
- Many interventions possible, none perfect
- Most of the world‘s malaria burden is in Africa
- Resource constrained context

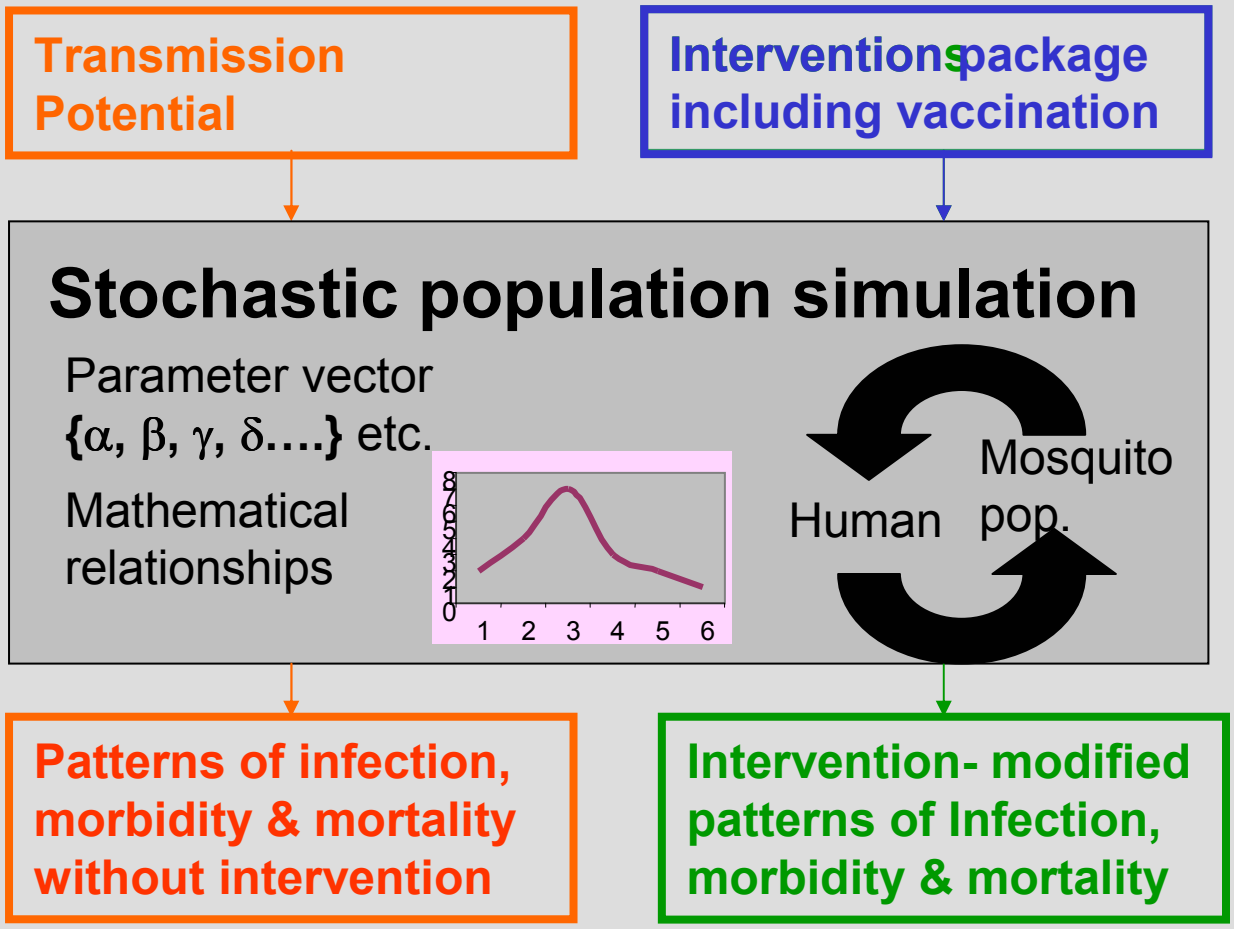


Origins of malariacontrol.net

- Initial project: Mathematical modeling of the impact of malaria vaccines on the clinical epidemiology and natural history of *P. falciparum* malaria (supported by Malaria Vaccine Initiative & GlaxoSmithKline from 2003-2005)
- Current extension to evaluate the likely impact of different control strategies singly and combined-
  - Vector control (mosquito nets, insecticide spraying of houses)
  - Different kinds of vaccines


Costs of interventions
Transmission Potential
Intervention package
Costs of disease
Patterns of infection, morbidity & mortality without intervention
Intervention- modified patterns of Infection, morbidity & mortality
Burden of disease without intervention
Intervention-modified burden of disease
Predicted intervention cost- effectiveness

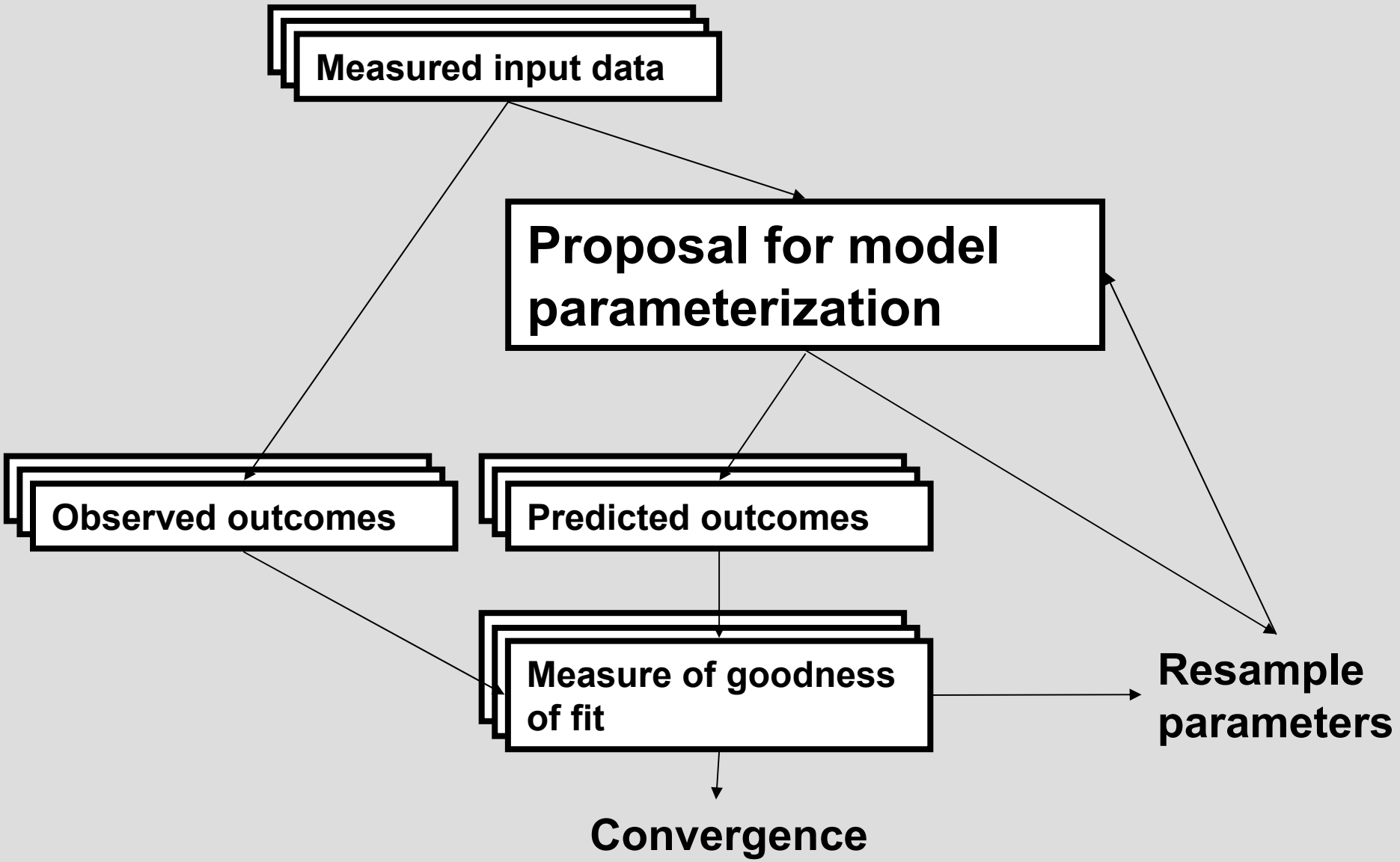
The black box



Modeling Approach

- Discrete time stochastic individual-based simulations
  - Hosts are characterized by a set of state variables (age, parasite densities, immune status variables, infectiousness)
- Empirical description of within-host asexual parasite densities
- Model for the effect of acquired immunity on parasite densities
- Models for transmission to the vector, for morbidity, and for mortality, as functions of parasite density
- Fit model to data from field studies
- Predict impact of control strategies by comparing simulated interventions with baseline scenarios

Estimating model parameters from field data

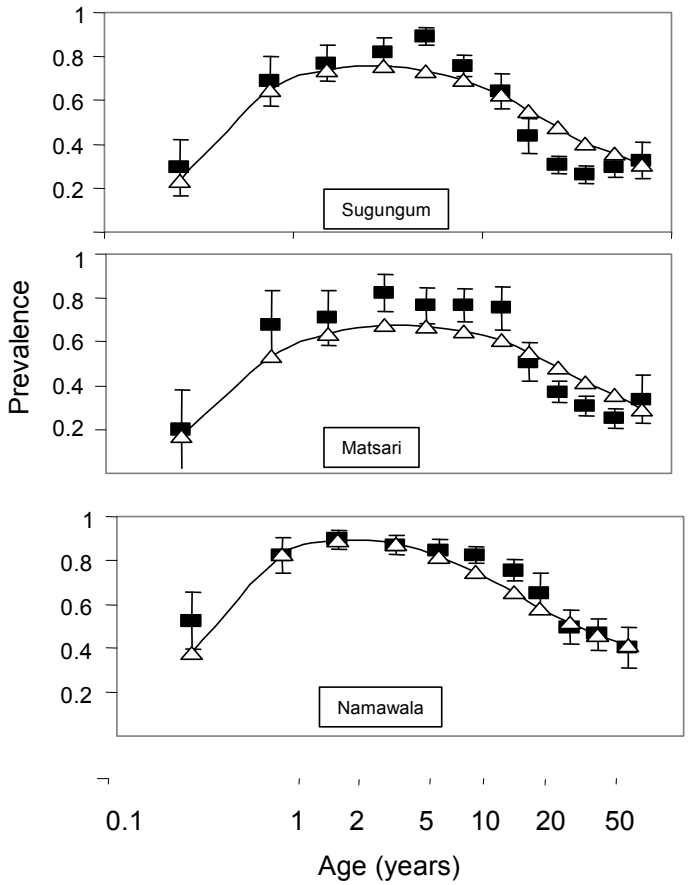


Optimization

- 61 datasets from field studies, different objectives*
  - Incidence of infection
  - Age-prevalence of parasitemia
  - Seasonality of parasitemia
  - Age-density of parasites
  - Age-incidence of clinical disease, hospitalisation and mortality

*all related to seasonal patterns of transmission

Age-prevalence of infection



Computation needs

- Individual-based approach
- Starting point: Immunological equilibrium
- Outcomes of interest are rare events
- Fitting of model parameters to field data
- Prediction for a range of scenarios
- Sensitivity analysis
- Model comparison

- Many millions of simulation runs, each in the order of hours
- Thousands of years of CPU time
- **A volunteer computing project**

malariacontrol.net: Suitability for Volunteer computing

- Independent parallelism
  - Divisible into parallel parts with few or no data dependencies
- Low data/compute ratio
  - Less than a gigabyte of data per day of CPU time
- Not dependent on short turnaround time
  - Several days per results, possibly resend a few times
- Public appeal

Port to BOINC

- Science application/Client
  - Reimplementation of some components
    - (Java-XML Databinding, NAG-libraries)
  - Communication with core client (BOINC-API)
  - Implementation of checkpointing
  - Screensaver graphics
- Project Server setup
  - Hardware donated by CERN/CUI Geneva
  - Hosting provided by CUI
  - Configuration and modification of project specific server components

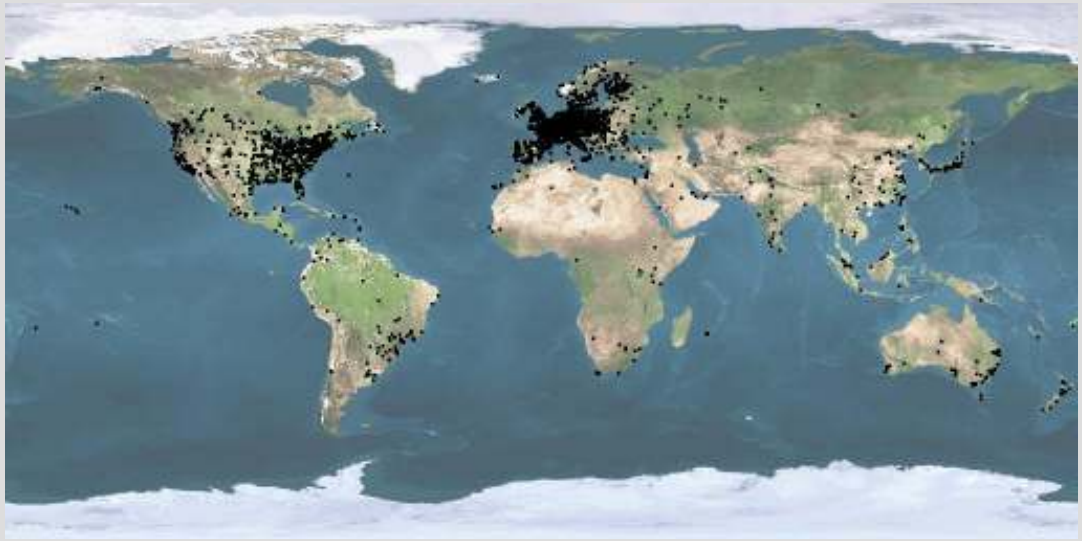
malariacontrol.net statistics

Volunteers: 9'000 total, 4500 active

- Sign up rate: up to 400 new users per day
- Currently 50-60 per day

Host PCs: 25,000 total, 15,000 active,

- 80% Windows, 20% Linux, Mac

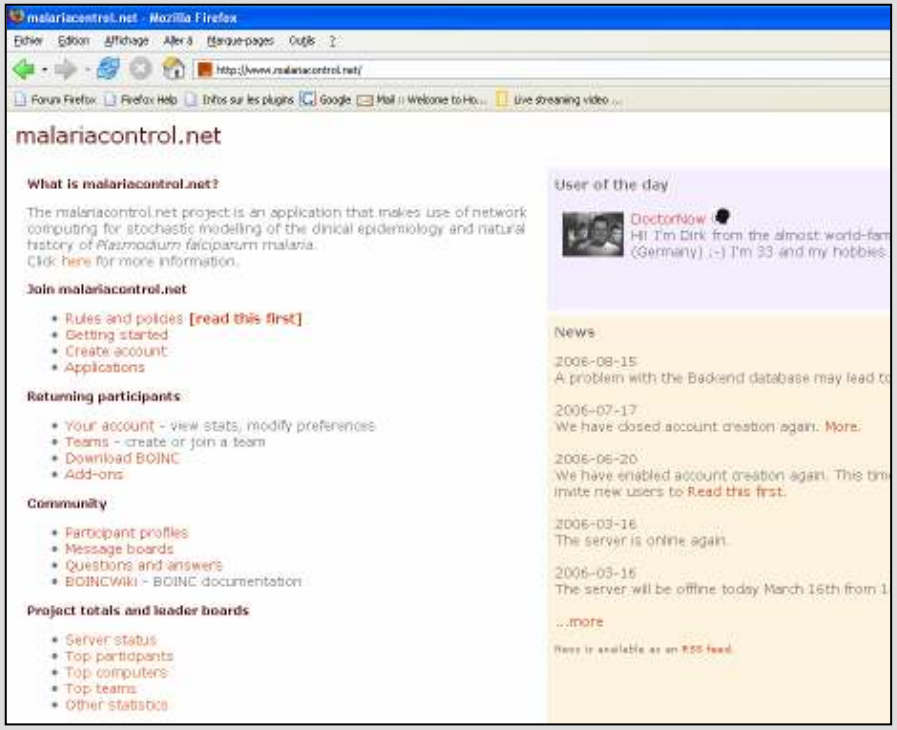


CPU power: 3.0 Teraflops

- equivalent to 1,000 CPU years/yr (midrange PCs)
- delivered to date 3,500 CPU years (Oct 07)

Simulations per day: 45,000

…+ huge public/press interest!



What is AFRICA@home?

Partnership set up to promote the use of volunteer computing for pressing **health and environmental issues** facing developing world.

- The goal of AFRICA@home is to involve **African students, scientists and institutions** in the development and running of these volunteer computing projects.
- The first application set up by AFRICA@home was malariacontrol.net**.**

Lastest development on Africa@Home projects

- **Currently two projects being developed:**
  - **AfricaMap**
  - **Docking** for Drug Discovery in Neglected Tropical Diseases

- **Developers: two selected Africans students from:**
  - Peter Amoako-Yirenkyi, Ghana
  - Eloi Appora-Gnékindy, Central African Republic

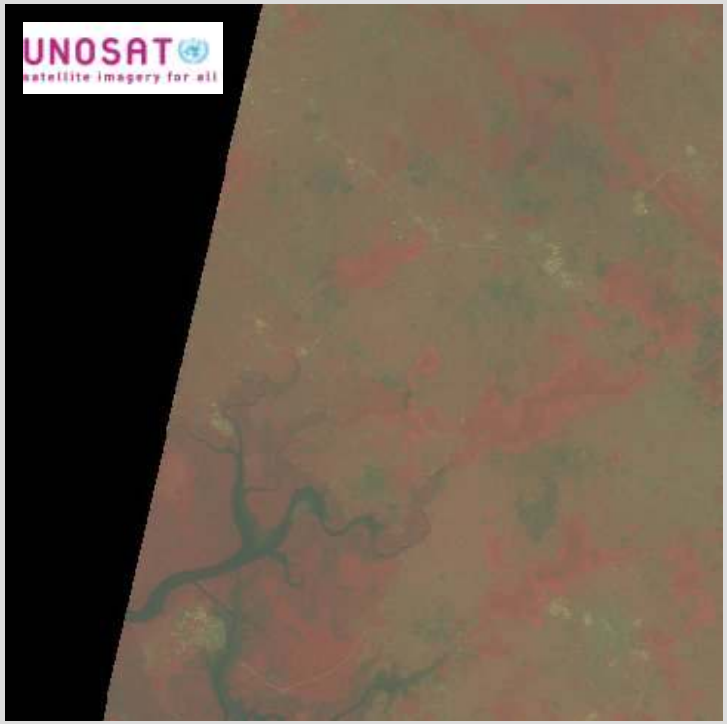
AfricaMap

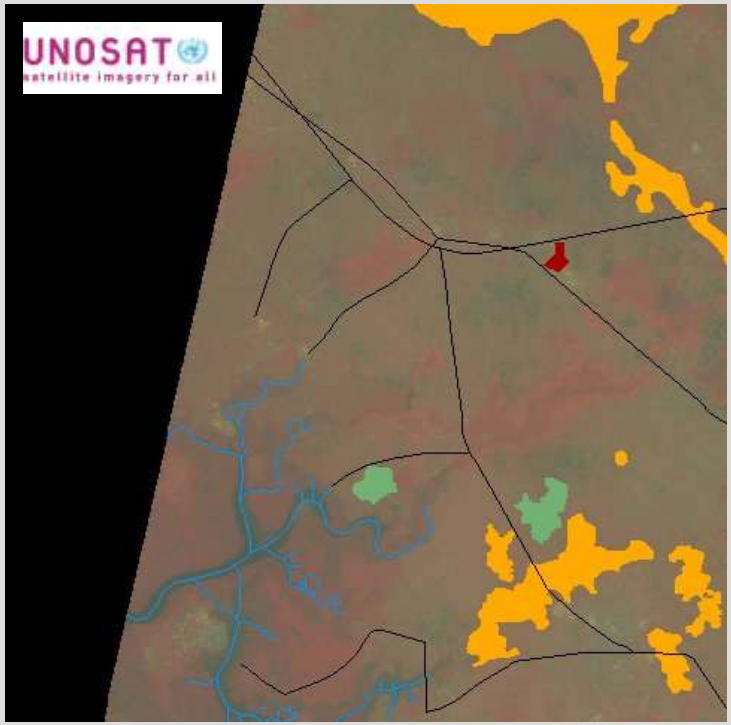
- Project partner: **UNOSAT**
- Goal: Implement a distributed framework for satellite images annotation.

UNOSAT *is a United Nations programme created to provide the international community and developing countries with enhanced access to satellite imagery and Geographic Information System (GIS) services. These tools are used mainly in humanitarian relief, disaster prevention and post crisis reconstruction.*

AfricaMap

People volunteer their skills in recognizing patterns from satellite imagery:
Roads, Houses, Landuse, Forest, Rivers, …





africa@home

- European Organization for Nuclear Research (CERN)
  - Ben Segal, Christian Soettrup, François Grey
- Centre Universitaire d'Informatique, Geneva (CUI)
  - Bastien Chopard, Christian Pellegrini
- International Conference Volunteers (ICV)
  - Viola Krebs
- Informaticiens sans Frontières (ISF)
  - Silvano de Gennaro
- UNOSAT
  - Ana Silva