

BioBITs

HiBB 2011 - Bordeaux - France

30th Aug 2011

On Parallelizing On-Line Statistics for Stochastic Biological Simulations

Marco Aldinucci

Mario Coppo, Ferruccio Damiani, Maurizio Drocco, Eva Sciacca, Salvatore Spinella, Angelo Troina

Computer Science Dept. - University of Torino - Italy

Massimo Torquati

Computer Science Dept. - University of Pisa - Italy

Outline

- Motivations & background
  - System biology and Calculus of Wrapped Sequences
  - Stochastic simulations on multi-core
- The CWC simulator with the on-line analysis tool: architecture
  - FastFlow: a lock-free pattern-based C++ template library
  - On the Architecture of Monte Carlo Simulators for Multicore
- Mono-stable and multi-stable systems, switches, oscillators
- On-line statistics, clustering, mining ...
  - Performance
  - Demo: switches and Lotka-Volterra (grass + sheep + wolves = ?)
- Conclusions




BioBITs

Modelling Complex Bio Systems

- A large effort to formally model complex systems is underway. Goal: developing a discipline for engineering
  - synthetic immune responses, virus diffusion, social behaviours, ...
- Two main approaches to study models
  - (Traditional) Ordinary Differential Equations (ODEs) + numerical solvers
  - Stochastic Process calculi + Monte Carlo Simulations (e.g. Gillespie ...)
    - Slower but able to (theoretically) model non-steady-state, non-average dynamic of systems




Computing Models for Systems Bio

- Lambda-calculus [Fontana & Buss, 1996];
- Petri nets [Matsuno et al., 2000];
- Process Calculi:
  - Biological π-calculus [Regev, Shapiro et al., 2001/2002]; BioAmbients [Regev et al. 2004]; Brane Calculi [Cardelli, 2005]; Beta-binders [Priami & Quaglia, 2005]; BioPEPA [Hillston et al., 2006];
- Rewrite Systems:
  - P-Systems [Paun, 1998]; κ-calculus [Danos & Laneve, 2003]; CLS [Barbuti et al. 2005]; Stochastic Bigraphs [Krivine et al., 2007]; **CWC** [Coppo et al. 2010];
- Statecharts [Harel et al., 2003];
- Hybrid Automata [Mishra et al. 2006]; ...



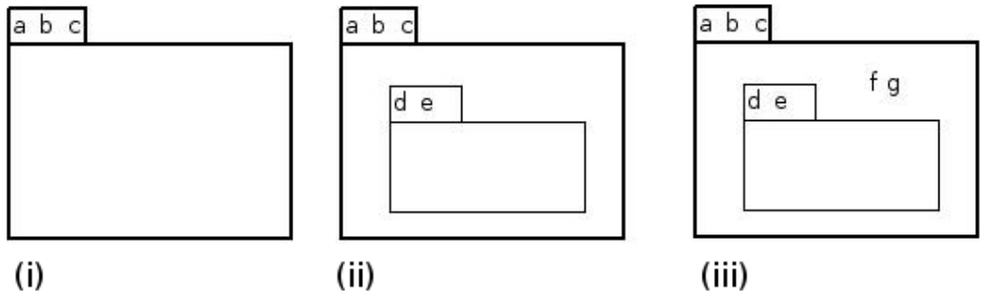

The Calculus of Wrapped Compartments (CWC)

A **term** is intended to represent a biological system. A *term* is built by means of the **compartment** constructor, $(- \rfloor -)$, from a set $\mathcal{E}$ of *atomic elements*, ranged over by a, b, c, d. A **simple term** is defined as:

$$t \quad ::= \quad a \;\Big|\; (\overline{a} \rfloor \overline{t})$$

We write $\overline{t}$ to denote a (possibly empty) multiset of simple terms $t_1 \ldots t_n$. Similarly, with $\overline{a}$ we denote a (possibly empty) multiset of atoms.

Examples of SCWC terms



- (i) represents $(a\ b\ c \rfloor \bullet)$;
- (ii) represents $(a\ b\ c \rfloor (d\ e \rfloor \bullet))$;
- (iii) represents $(a\ b\ c \rfloor (d\ e \rfloor \bullet)\ f\ g)$.

Dynamics of SCWC

Rewrite rules are defined as pairs of terms, in which the left term characterizes the portion of the system in which the event modelled by the rule can occur, and the right one describes how that portion of the system is changed by the event.

Biomolecular Event	**Examples of CWC Rewrite Rules**
State change	$a \mapsto b$
Complexation	$a\ b \mapsto c$
Catalyzed membrane crossing	$a\ (b\ x \rfloor y) \mapsto (b\ x \rfloor a\ y)$ $(b\ x \rfloor a\ y) \mapsto a\ (b\ x \rfloor y)$

Stochastic Rules

Rules are decorated with a **rate** (speed of the reaction).

A **Stochastic Rewrite Rule**, R, is denoted by $P \overset{k}{\mapsto} P'$.

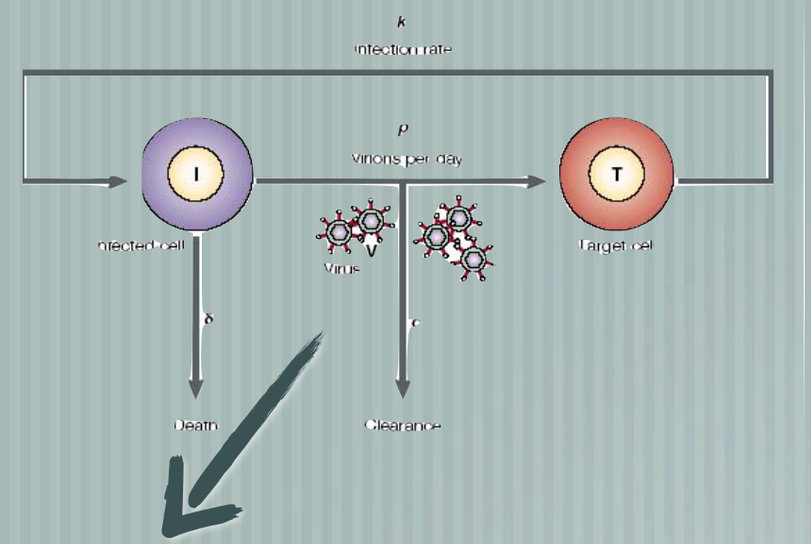
The stochastic semantics is given by transitions between terms labeled with the rule applied, R, and a transition rate depending on the rate of rule R:

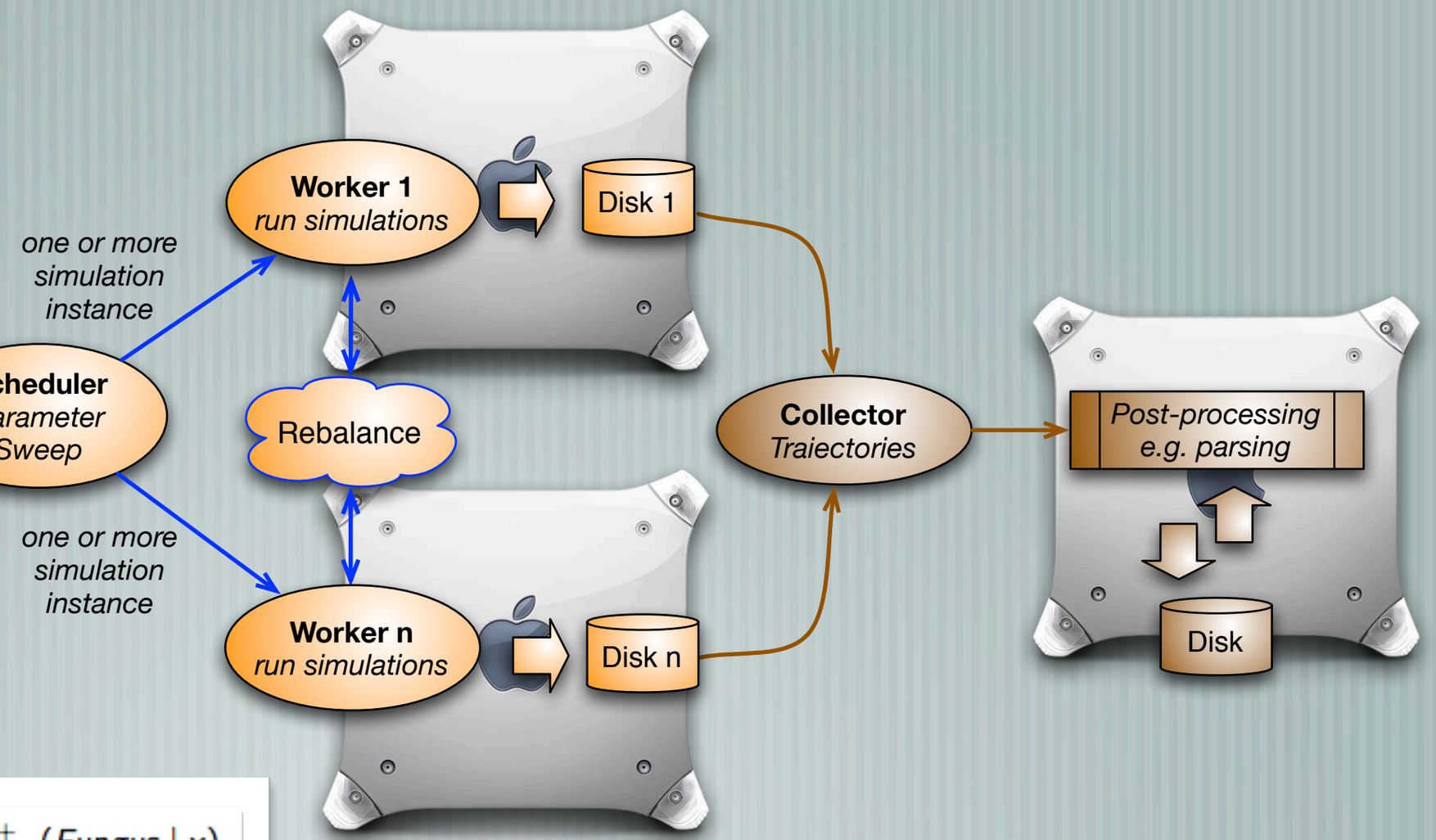
$$\overline{t} \overset{R, k \times p}{\longrightarrow} \overline{t'}$$

where R is $P \overset{k}{\mapsto} P'$, and p is the number of different ways in which the pattern P may match $\overline{t}$ ($\overline{t} = C[P\sigma]$) and such that $\overline{t'} = C[P'\sigma]$ for some context C and variable instantiation σ.

http://mc-fastflow.sourceforge.net

MonteCarlo sim: distributed solution



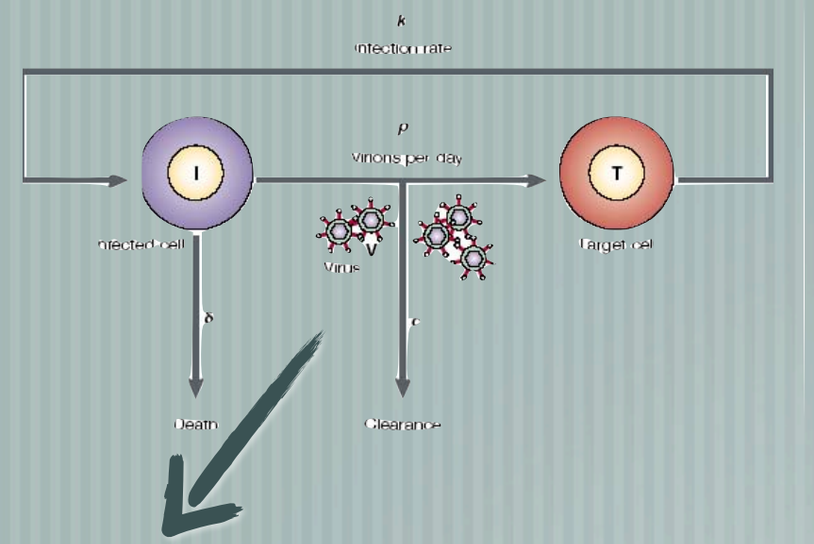


$$NH_3 \underset{k_2}{\overset{k_1}{\rightleftharpoons}} NH_4^+$$

$$(Fungus \rfloor NH_4^+ \;\; x) \overset{k_f}{\mapsto} NH_4^+ \;\; (Fungus \rfloor x)$$

BioBITs

MonteCarlo sim: distributed solution



In BioSims the whole trajectory is needed, easily GB of data, e.g. 5.6 GB for HIV 4-year for a single run (*)

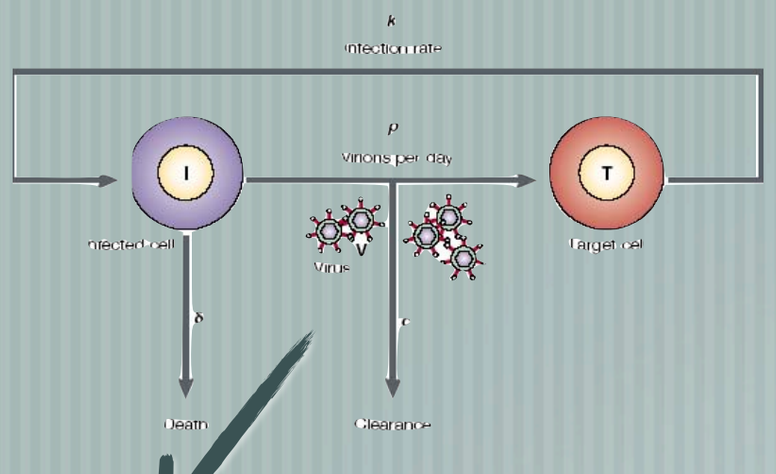
The whole dataset, TB of data, should be re-read to extract statistical estimators

(*) M. Aldinucci, A. Bracciali, P. Liò, A. Sorathiya, and M. Torquati. StochKit-FF: Efficient systems biology on multicore architectures. In Euro-Par Workshops 2010, LNCS, Ischia, Italy, Sept. 2010. Springer.

BioBITs

http://mc-fastflow.sourceforge.net

MonteCarlo sim: on multicore?



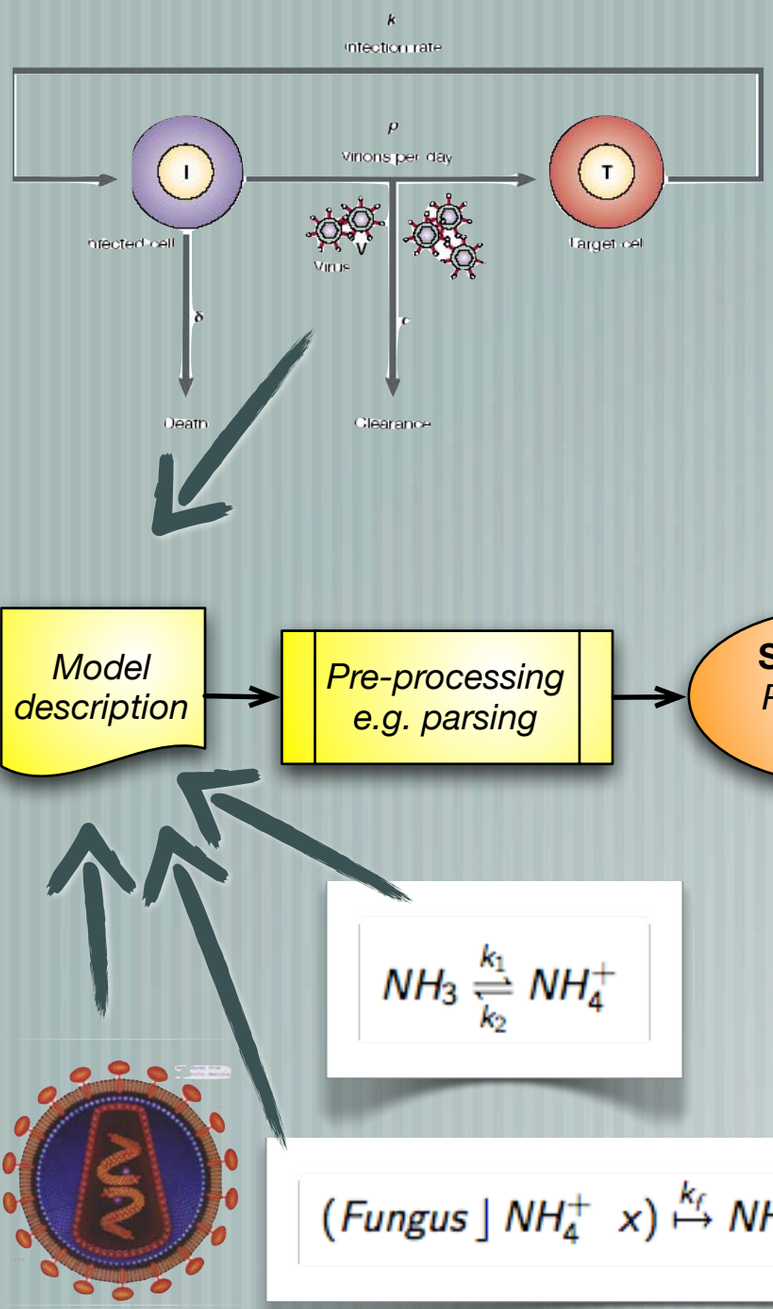
$$NH_3 \underset{k_2}{\overset{k_1}{\rightleftharpoons}} NH_4^+$$

$$(Fungus \rfloor NH_4^+ \;\; x) \overset{k_f}{\mapsto} NH_4^+ \;\; (Fungus \rfloor x)$$

BioBITs

http://mc-fastflow.sourceforge.net

MonteCarlo sim: on multicore?



Now the issue become a real problem

Bottlenecks: disk and memory

Post-processing: not pipelined and often not parallel at all

$$NH_3 \underset{k_2}{\overset{k_1}{\rightleftharpoons}} NH_4^+$$

$$(Fungus \rfloor NH_4^+ \ x) \overset{k_f}{\mapsto} NH_4^+ \ (Fungus \rfloor x)$$

BioBITs

From Distributed to Multicore

- MultiCarlo sims for Bio are I/O-bound
  - The whole trajectory is needed
  - Sampling reduce I/O traffic but worsen precision and analysis of "strange" dynamics (spikes, diversion from average, etc.), which observation motivates stochastic analysis (ODEs)
- Data analysis is also I/O-bound
  - if approached is a "post-processing" fashion, data should be retrieved from the disks
- The porting of distributed solution "as is" on multicore is going insist on weak points of multicore architectures
  - Memory wall, I/O, disk
  - SIMD/GPGPUs do not change the analysis substantially




BioBITs

Manage large data set on multicore

- Biological data is typically huge
  - not only simulators but also data from DB/web, analysis instruments, ...
  - raw data is often obscure and analysis can be very expensive
- Rationale
  - Manage data as stream, compute everything online
    - included statistics and data analisys
  - Establish fast data paths across cores
  - Avoid low-level concurrency management
    - Portability, performance, portability of performance, maintenance, porting from sequential

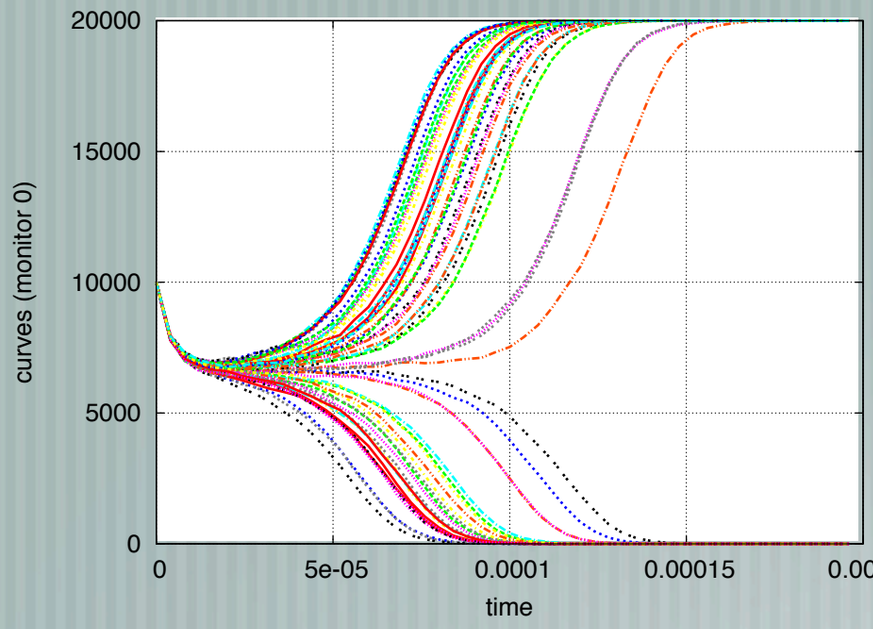


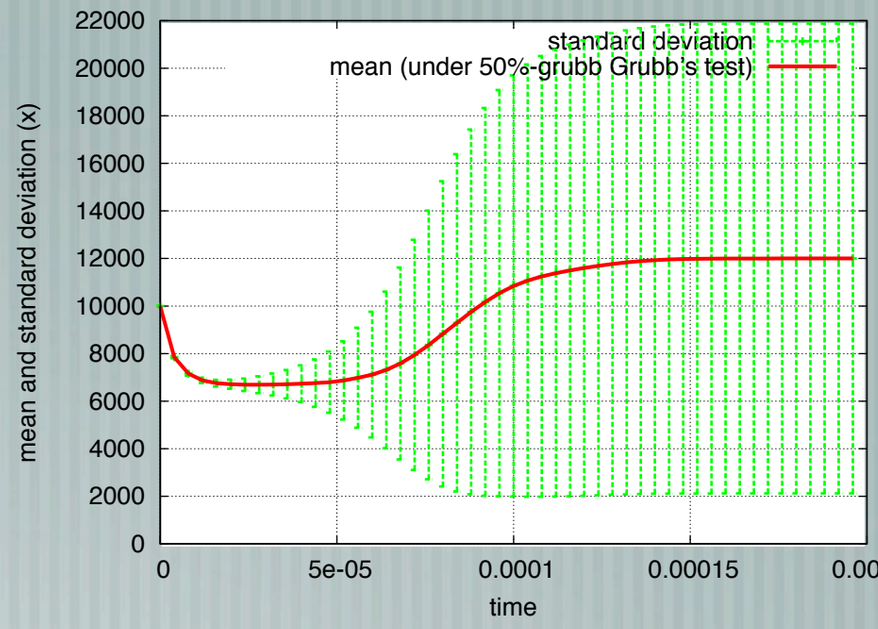

Which are the most interesting biological systems for (high-perf) stocastic simulations?

- Stable system
  - Well-understood from biologists
  - Can be also addressed with ODEs (Gillespie)
    - ODE solvers typically faster (and then hybrid approaches, ...)
- Multi-stable and unstable systems
  - Can be studied with stochastic simutions
  - Cannot be easily addressed with ODEs
  - Can be used to model interesting phenomena
    - deviant behaviours, oscillating systems, genetic switches, etc.




Multi-stable systems and data analisys





- Classic statistical estimators (e.g. mean) almost useless
- More sophisticated data classification is needed
  - classification, clustering, mining, etc ...
  - the useful mining tool is often unknown ex-ante




BioBITs

Parallel simulation with parallel analysis

- Parallel simulation
  - exploiting independence among different simulation
  - run in a lock-simulation-step fashion in order to push forward an almost aligned simulation front
- A battery of analysis tools
  - both run on successive (but collective) data windows
  - exploiting independent or correlated analysis tools
    - synchronisation should enforce possible data dependencies

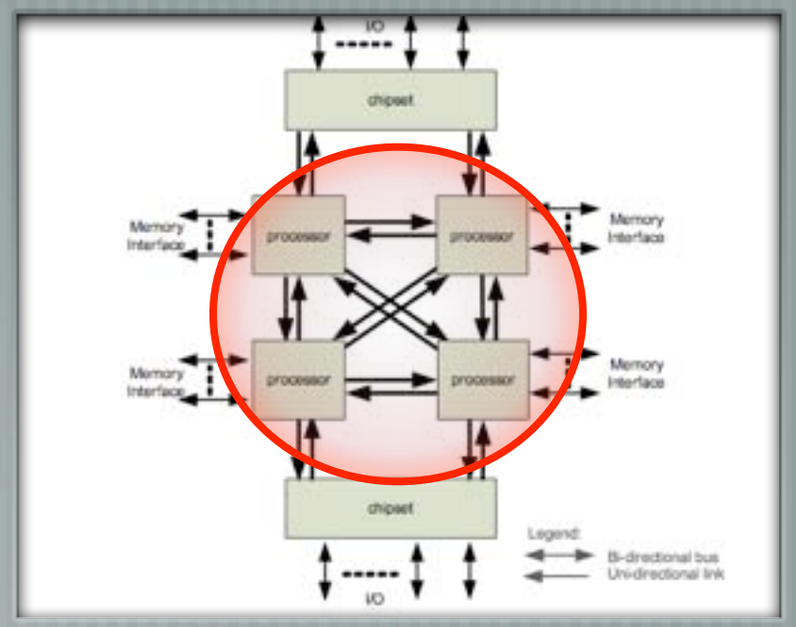



BioBITs

The CWC simulator with the on-line analysis tool: architecture

This and next generation Multi-cores

- Are programmed at "concurrent assembler" level
  - Complex, not portable, not efficient
- Exploit cache coherence
  - Lock/Memory-Fences are expensive
  - Will worsen with core count
  - Atomic ops do not solve the problem (still fences)
- Fine-grained parallelism is the problem
  - I/O bound problems, High-throughput, Streaming, Irregular DP problems



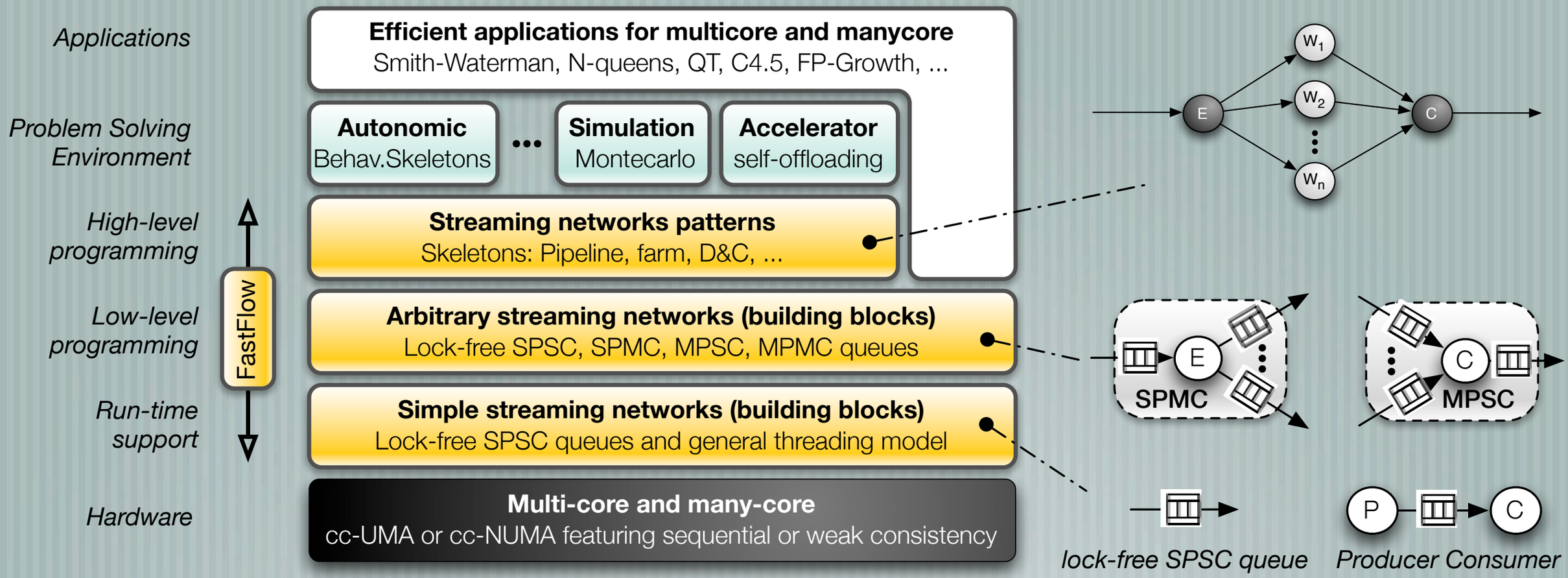
[2009] i7 QuickPath
(MESI-F Directory Coherence)




BioBITs

http://mc-fastflow.sourceforge.net

FastFlow: easy streaming in C++

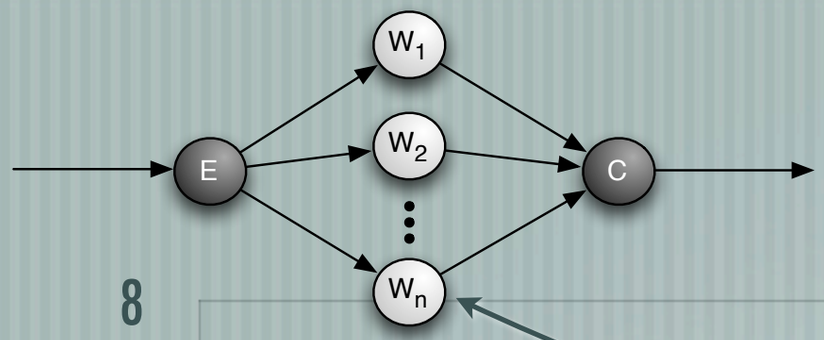


High-level programming

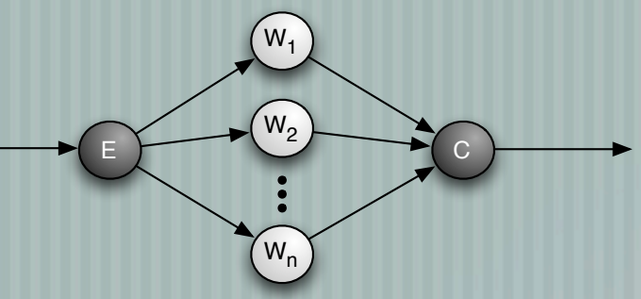
- Implemented on top of lock-free/fence-free non-blocking synchronisations
- C++ STL-like implementation

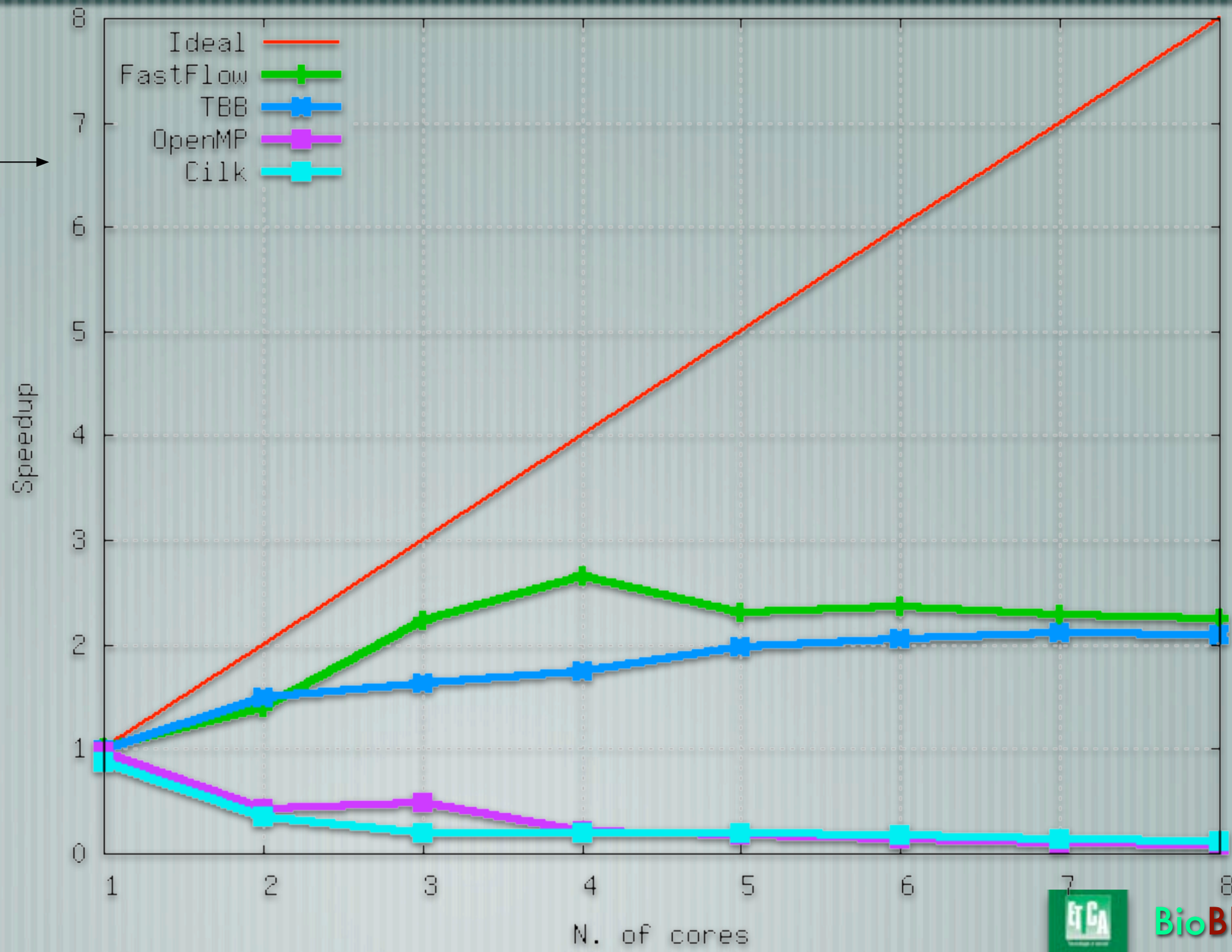
BioBITs

E.g. farm with POSIX lock/unlock



E.g. farm with TBB, OpenMP, Cilk





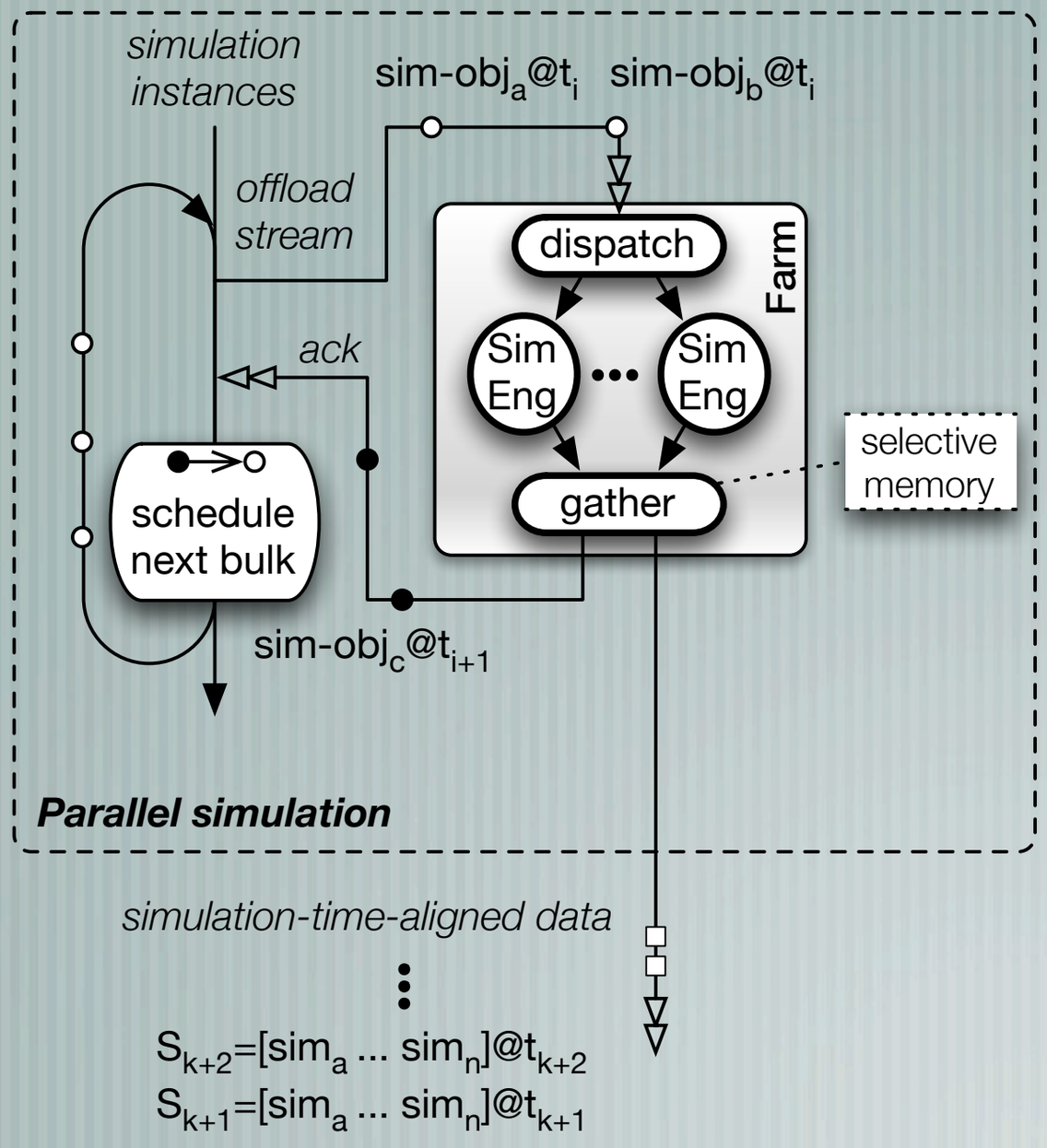
BioBITs

Why efficiency at fine-grain is important?

- It is enabling feature for the scalability of irregular, recursive apps; it reduces the programming and tuning effort
  - Gillespie's sim need about 50 instructions per iteration
- Other issues for Monte Carlo simulators
  - Data stream as a first-class concept (cannot store the whole dataset)
  - Assisted porting methodology (easy migration existing codes)
  - Cache-friendly synchronisation for data streams (performance)
  - Load balancing of irregular workloads (performance, portability)

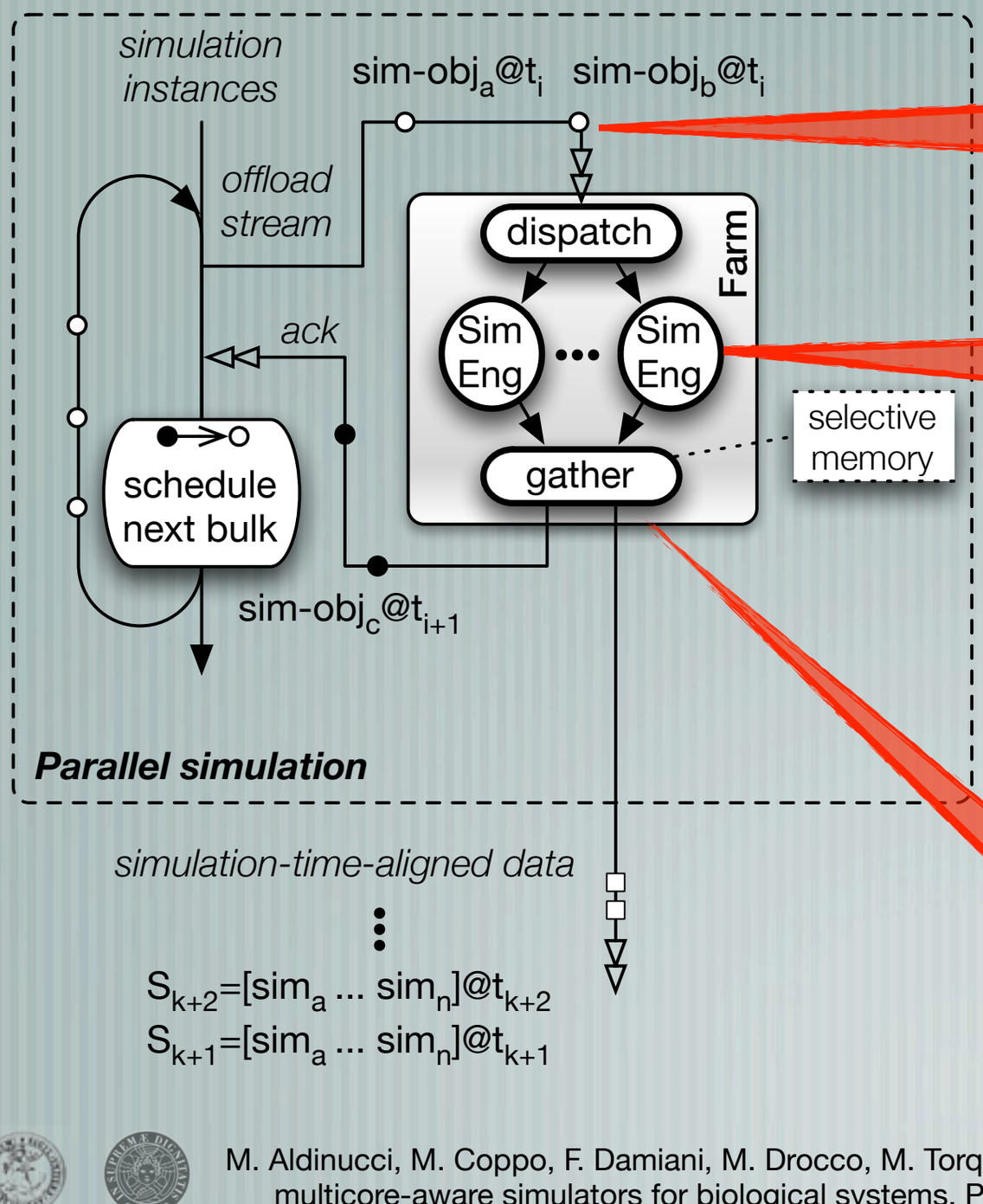



Architecture: Simulator



M. Aldinucci, M. Coppo, F. Damiani, M. Drocco, M. Torquati, and A. Troina. On designing multicore-aware simulators for biological systems. PDP 2011, Feb. 2011. IEEE.

Architecture: Simulator



Tokens are pointers to simution objects in the shared mem.
They are dispached to different Sim Eng according to a user defined policy

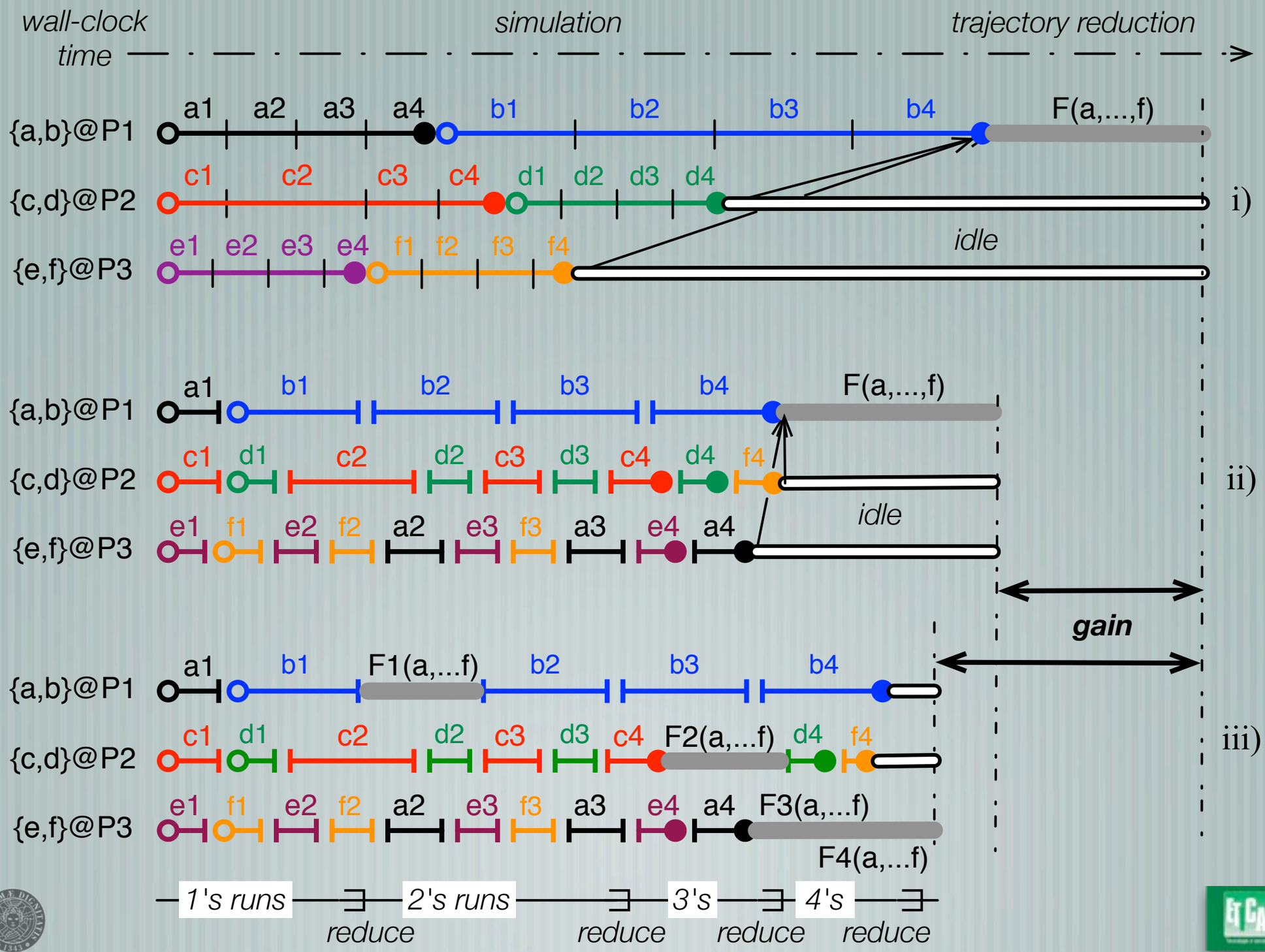
SimEng push forward the simulation a given amount of time.
Sim Eng has the same code of seq simulator.
The library manage thread creation, synchronizations, etc.

Simulations are aligned in time, completed simulation are outputted

M. Aldinucci, M. Coppo, F. Damiani, M. Drocco, M. Torquati, and A. Troina. On designing multicore-aware simulators for biological systems. PDP 2011, Feb. 2011. IEEE.

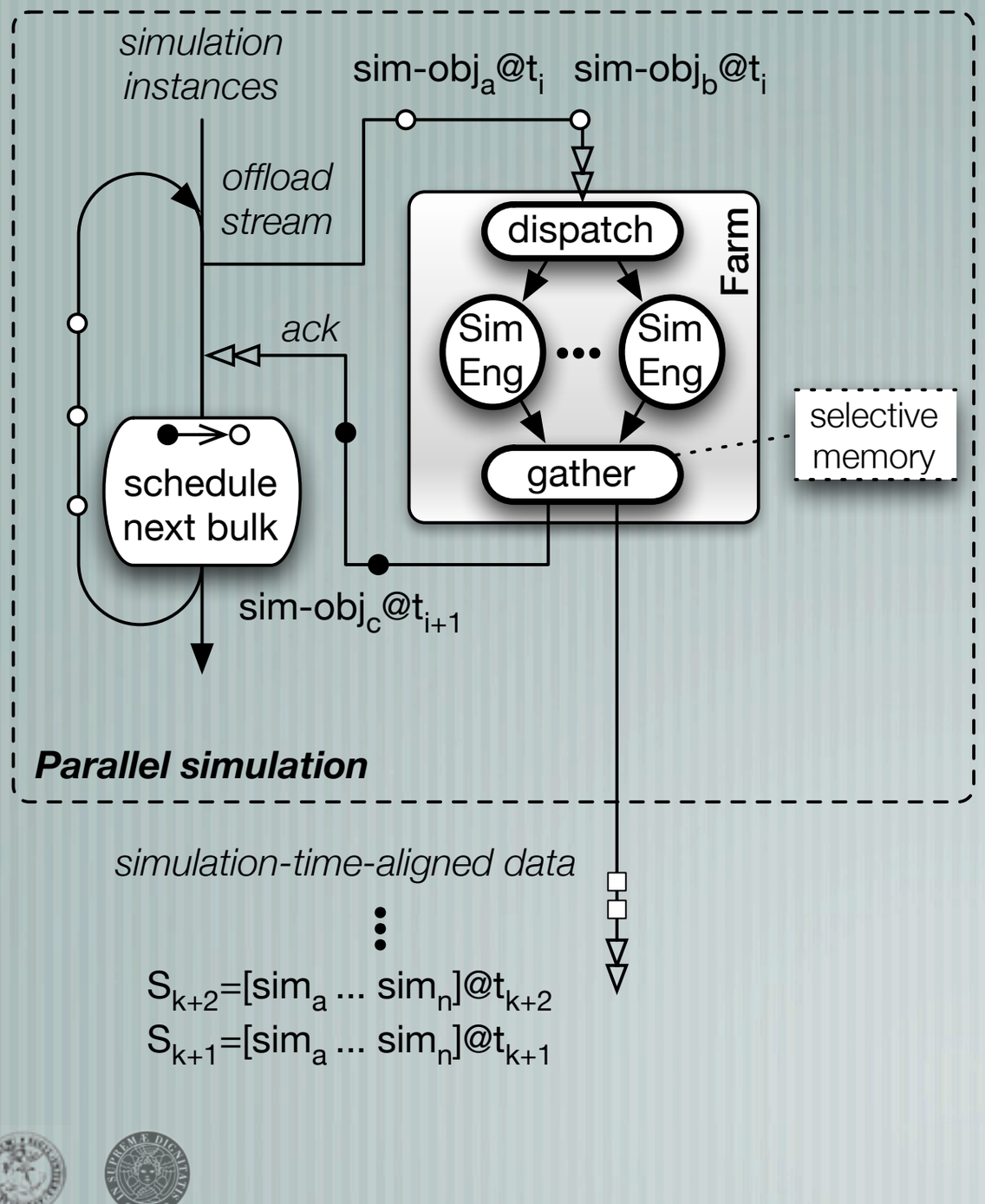
BioBITs

Three scheduling policies: i) Round Robin ii) Auto-balancing iii) Auto-balancing with pipelined reduction

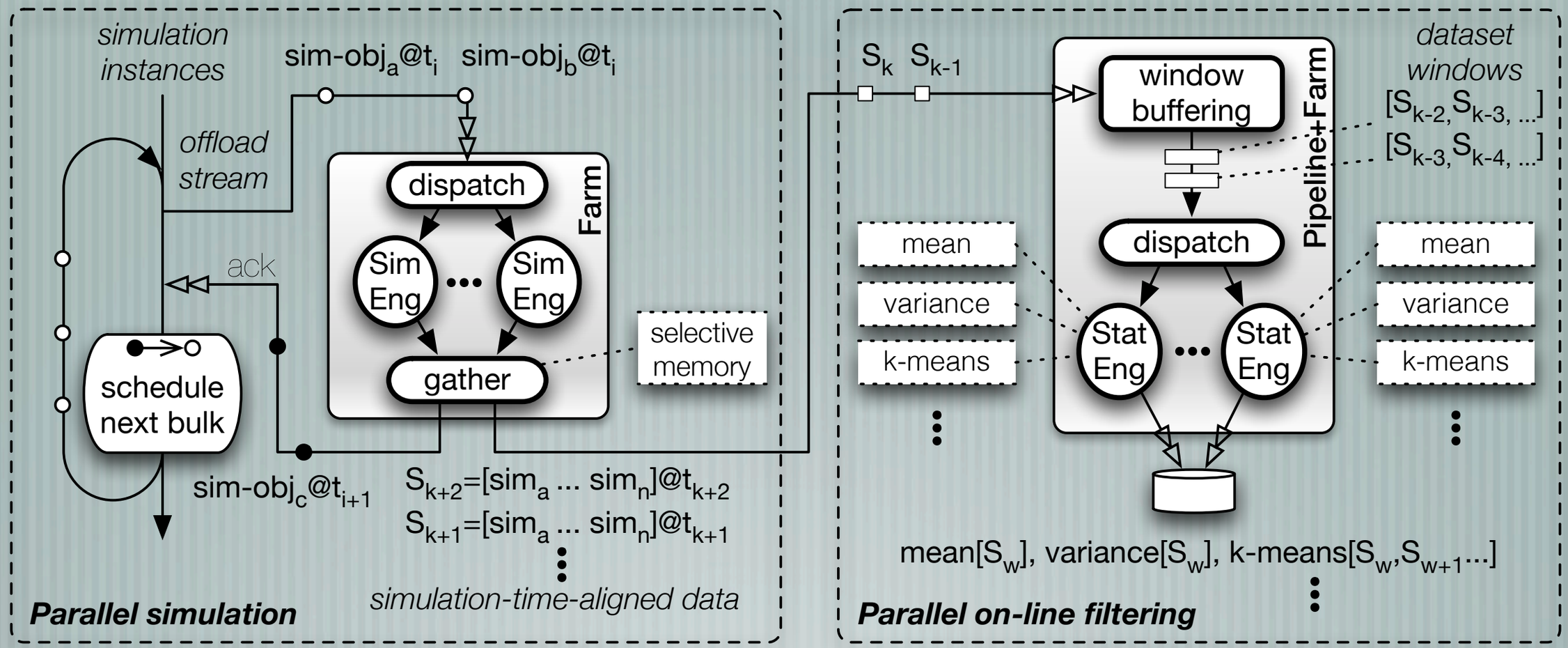


BioBITs

Architecture: Simulator

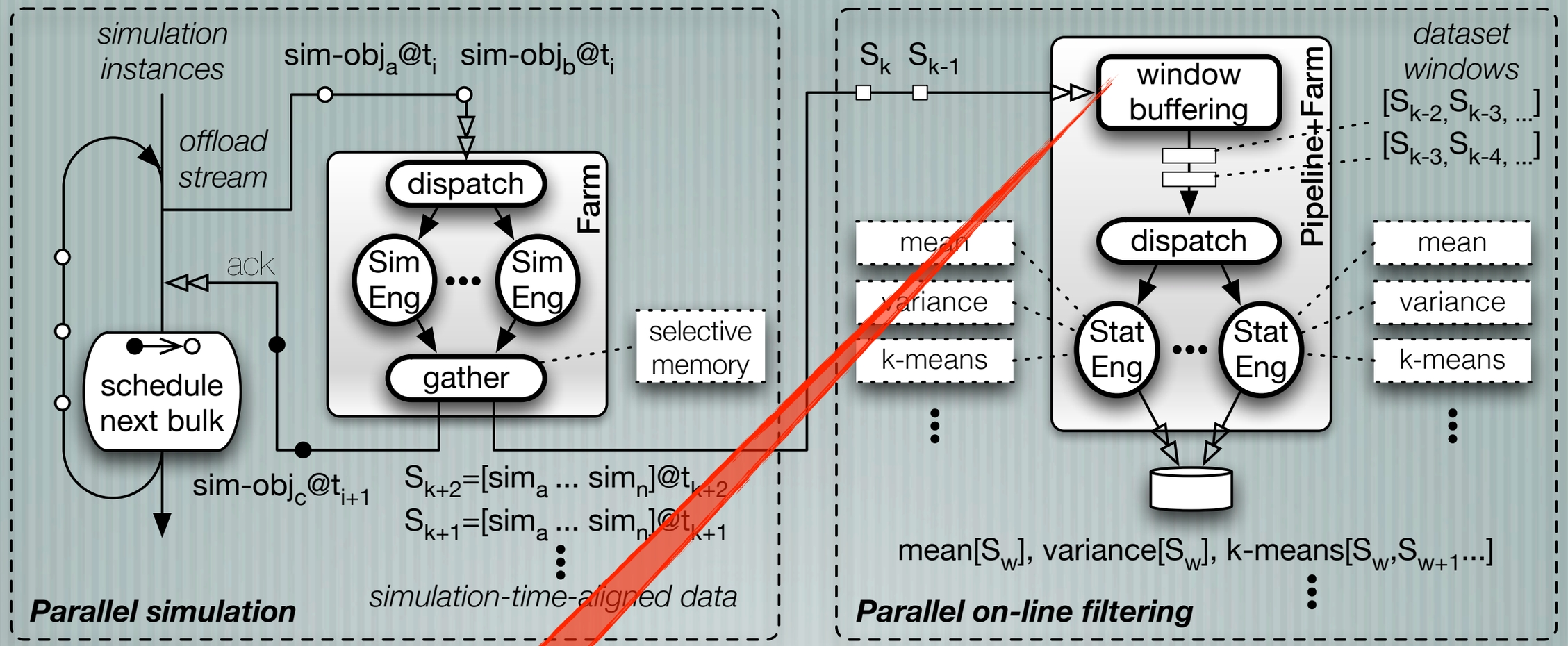


Architecture: Simulator + Online Stats



BioBITs

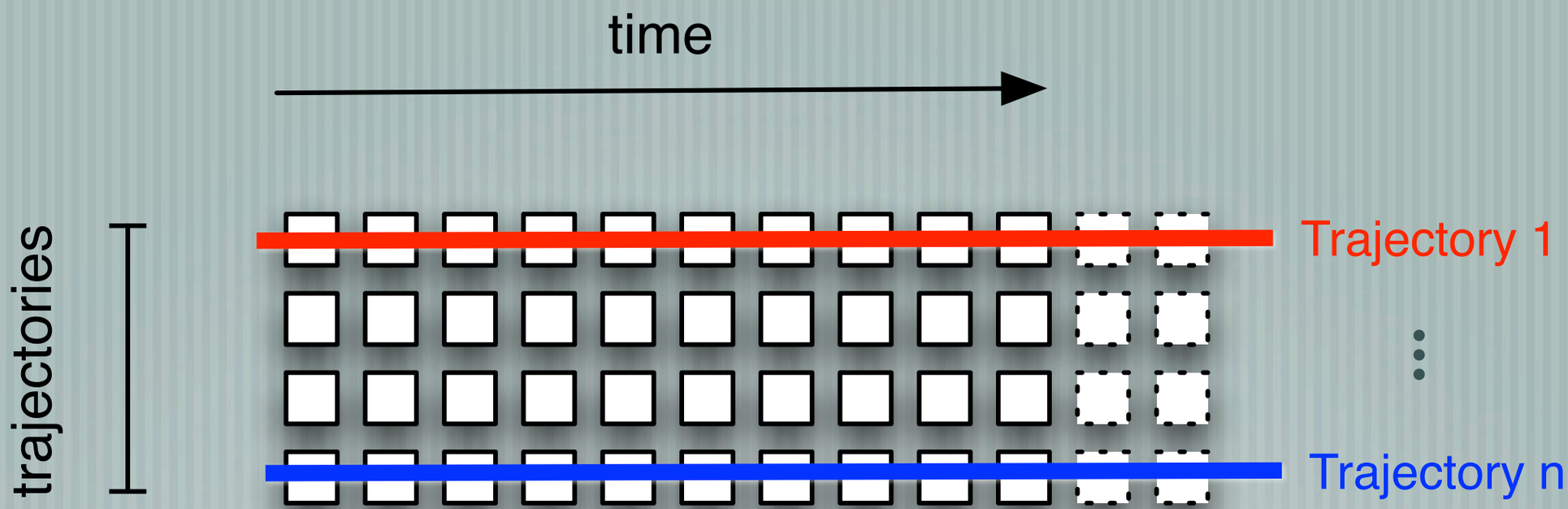
Architecture: Simulator + Online Stats



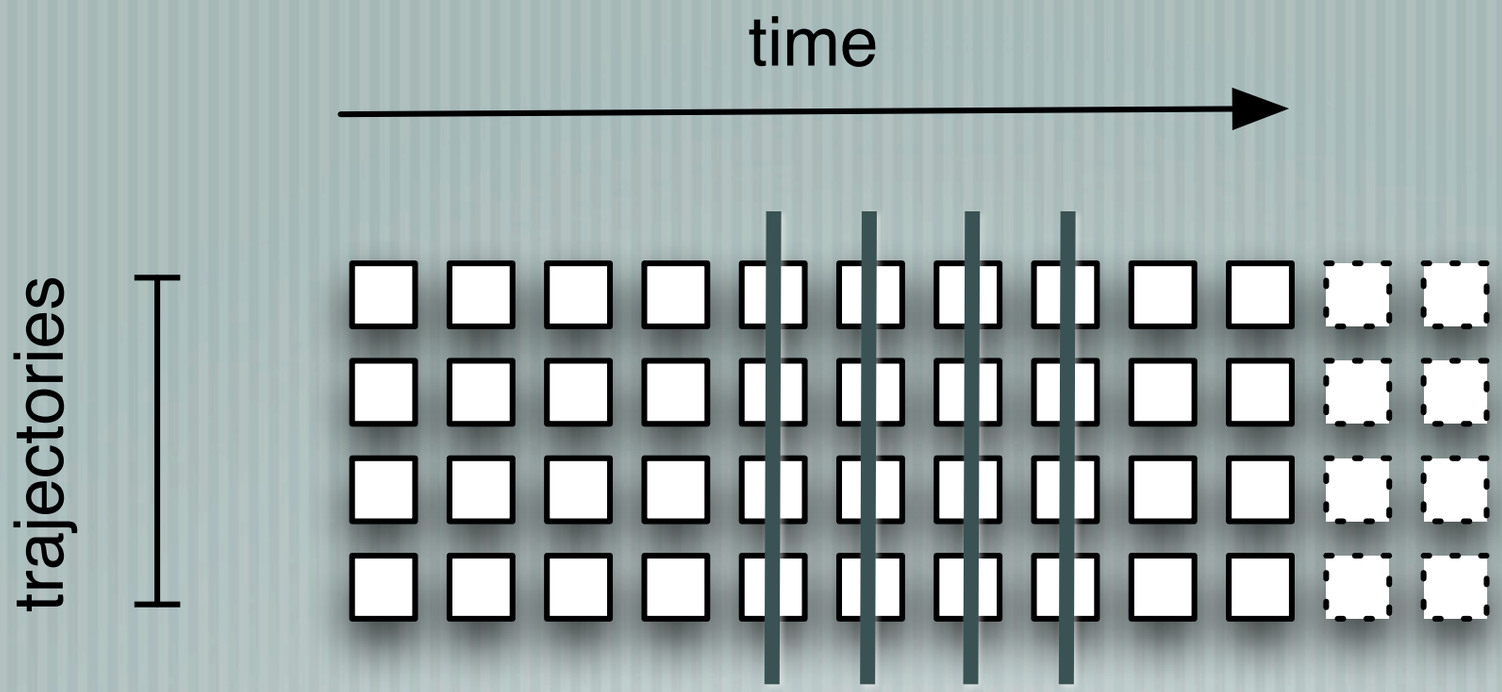
A window of simulation-time trajectories are buffered. The window slides forward

BioBITs

Sliding windows and running stats

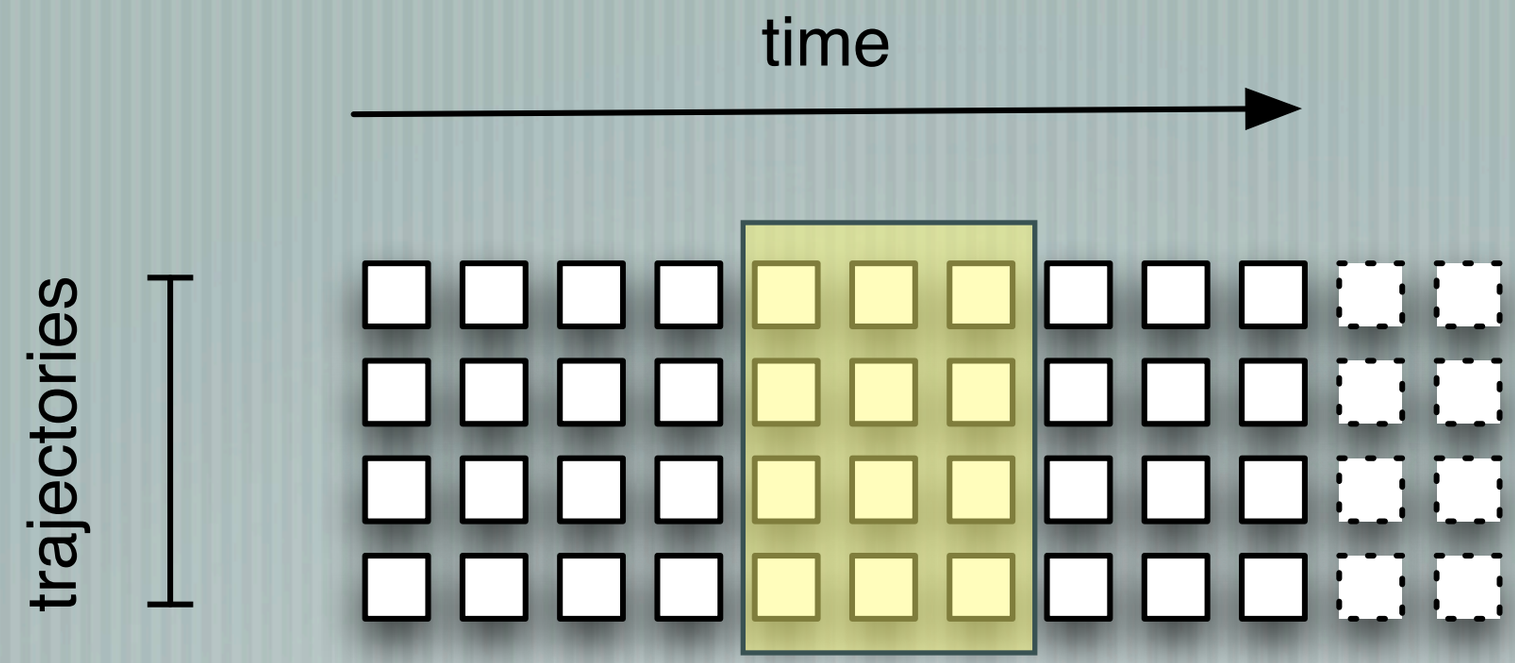


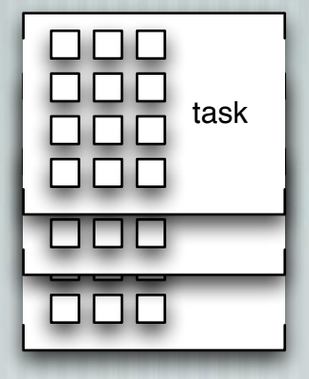
Sliding windows and running stats



simulation-time aligned trajectories

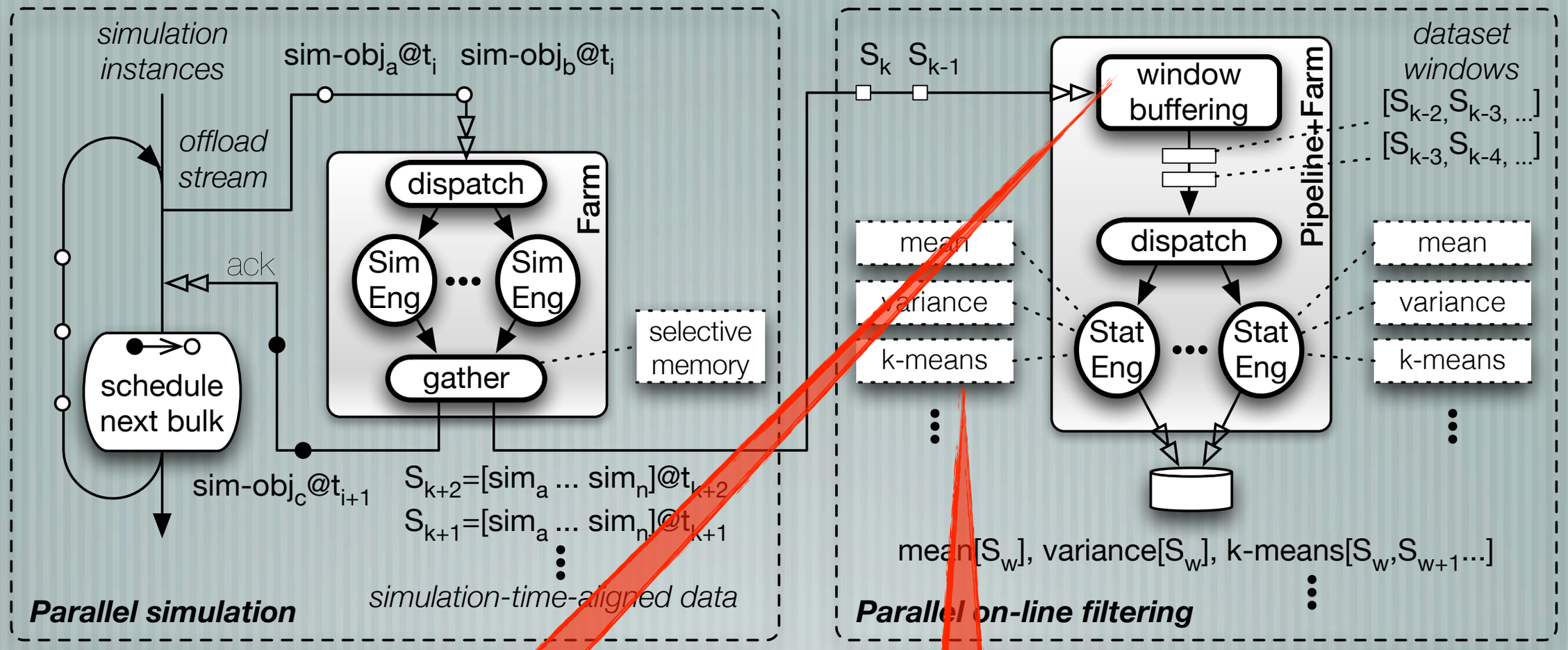
Sliding windows and running stats





dispatch

Architecture: Simulator + Online Stats



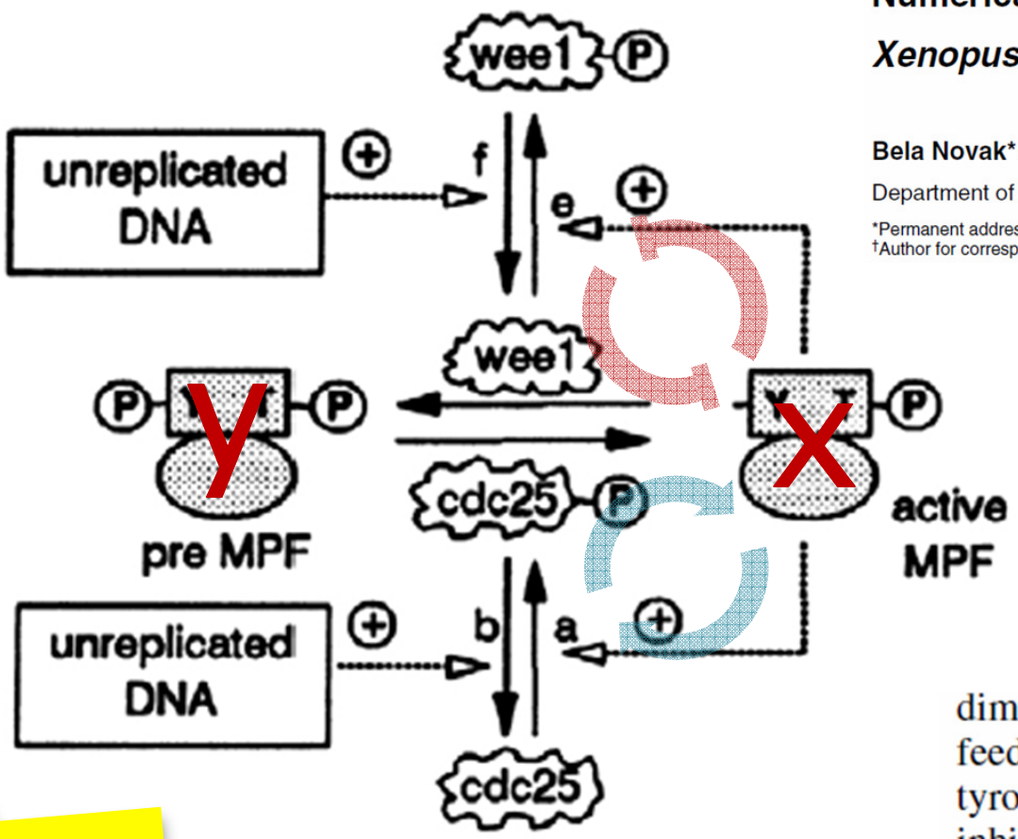
A window of simulation-time trajectories are buffered. The window slides forward

A user-defined battery of statistic or mining tools are run in parallel on different windows

BioBITs

Multi-stable systems

Formalising the cell cycle switch



Journal of Cell Science 106, 1153-1168 (1993)
Printed in Great Britain © The Company of Biologists Limited 1993

Numerical analysis of a comprehensive model of M-phase control in *Xenopus* oocyte extracts and intact embryos

Bela Novak* and John J. Tyson†

Department of Biology, Virginia Polytechnic Institute and State University, Blacksburg, Virginia 24060-0406, USA

*Permanent address: Department of Agricultural Chemical Technology, Technical University of Budapest, 1521 Budapest Gellert Ter 4, Hungary
†Author for correspondence

dimers is left off the diagram to keep it simple.) (B) Positive feedback loops. Active MPF stimulates its own production from tyrosine-phosphorylated dimers by activating Cdc25 and inhibiting Wee1. We suspect that these signals are indirect, but intermediary enzymes are unknown and we ignore them in this paper. The signals from active MPF to Wee1 and Cdc25 generate an autocatalytic instability in the control system. We indicate also an 'external' signal from unreplicated DNA to Wee1 and Cdc25, which can be used to control the efficacy of the positive feedback loops. The letters a, b, e and f are used to label the rate constants for these reactions in Fig. 2. (C) Negative feedback loop. Active

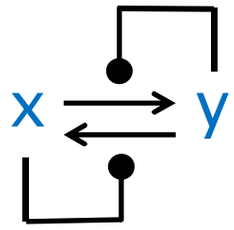
from Luca Cardelli's talk

On Switches and Oscillators Program Equivalence in Biology?

http://lucacardelli.name

Direct competition: unstable switch

- x catalyzes the transformation of y into x
- y catalyzes the transformation of x into y



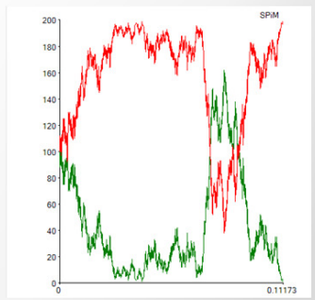
$$x + y \rightarrow x + x$$
$$y + x \rightarrow y + y$$

from Luca Cardelli's talk

On Switches and Oscillators Program Equivalence in Biology?

http://lucacardelli.name

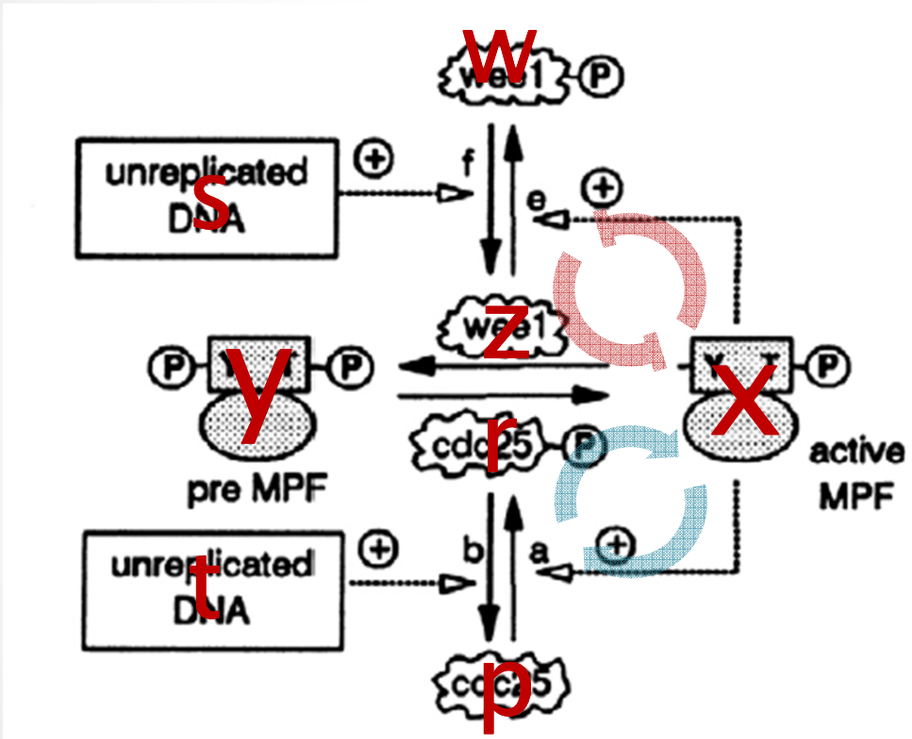
- This system is bistable, but
  - Convergence to a stable state is slow (a random walk).
  - *Any* perturbation of a stable state can initiate a random walk to the other stable state.
  - With 100 molecules of x and y, convergence is quick, but with 10000 molecules, even at the same concentration (adjusting the rate) you will wait for a long time.

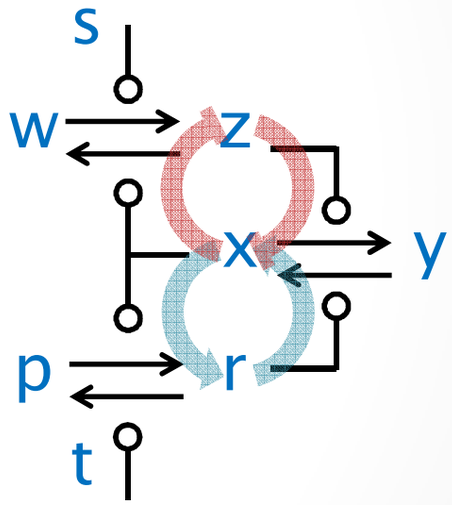


CWC syntax

$$\top : a\ c \xmapsto{10} a\ a \qquad \top : c\ a \xmapsto{10} c\ c$$

... after a number of transformations: a stable switch faithfully modelling cell switch





from Luca Cardelli's talk

On Switches and Oscillators Program Equivalence in Biology?

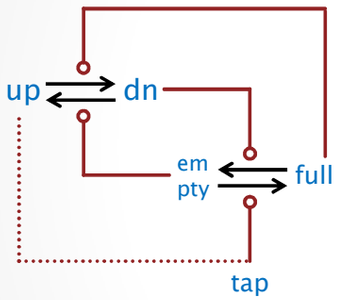
http://lucacardelli.name

CWC syntax

$$\top : a\ c \xmapsto{10} c\ b \qquad \top : c\ a \xmapsto{10} a\ b \qquad \top : b\ a \xmapsto{10} a\ a \qquad \top : b\ c \xmapsto{10} c\ c$$

The Shishi Odoshi

- A Japanese scarecrow (scare-deer)
  - Used by Bela Novak to illustrate the cell cycle switch.



http://www.youtube.com/watch?v=Vbv

empty + tap → tap + full
up + full → full + dn
full + dn → dn + empty
dn + empty → empty + up

To make it into a full trammel (do
could make the up position mecha
tap (i.e. take up = tap)

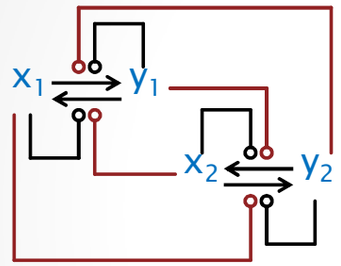
from Luca Cardelli's talk

On Switches and Oscillators Program Equivalence in Biology?

http://lucacardelli.name

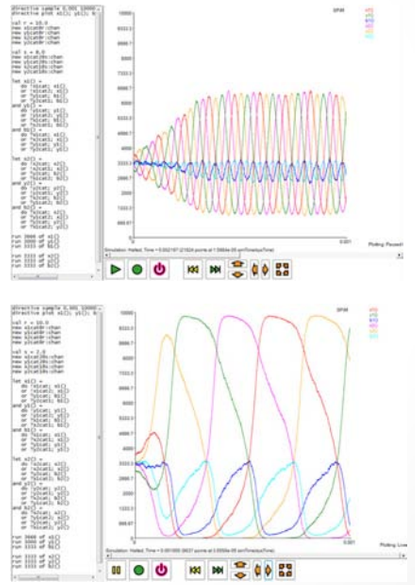
The 2AM Limit-Cycle Oscillator

- Two AM switches in a Trammel pattern



d to be slower
e black reactions,
illation is robust.
) vs. 10 and 1 vs.
8 vs 10.0 top, and

illators in the
ical rate ranges.)

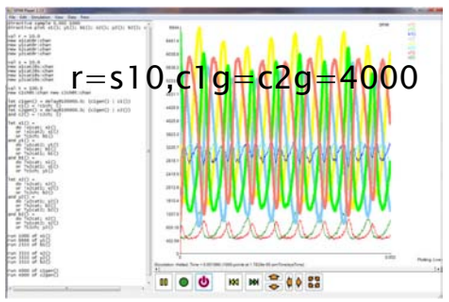


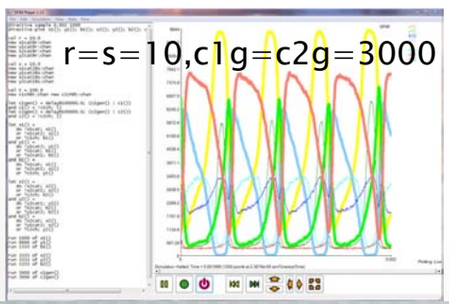
```
directive sample 0.001 10000
directive plot x1(); y1(); b1(); x2();
y2(); b2()

val r = 10.0
new x1cat@r:chan
new y1cat@r:chan
new x2cat@r:chan
new y2cat@r:chan

val s = 8.0
new x1cat2@s:chan
new y1cat2@s:chan
new x2cat1@s:chan
new y2cat1@s:chan

let x1() =
  do !x1cat; x1()
  or !x1cat2; x1()
  or ?y1cat; b1()
  or ?y2cat1; b1()
and y1() =
  do !y1cat; y1()
  or !y1cat2; y1()
  or ?x1cat; b1()
  or ?x2cat1; b1()
and b1() =
  do ?x1cat; x1()
  or ?x2cat1; x1()
  or ?y1cat; y1()
  or ?y2cat1; y1()

let x2() =
  do !x2cat; x2()
  or !x2cat1; x2()
  or ?y2cat; b2()
  or ?x1cat2; b2()
and y2() =
  do !y2cat; y2()
  or !y2cat1; y2()
  or ?x2cat; b2()
  or ?y1cat2; b2()
and b2() =
  do ?x2cat; x2()
  or ?y1cat2; x2()
  or ?y2cat; y2()
  or ?x1cat2; y2()

run 3666 of x1()
run 3000 of y1()
run 3333 of b1()

run 3333 of x2()
run 3333 of y2()
run 3333 of b2()
```

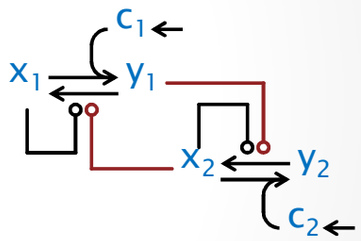
Influx Oscillators

- Similar but:
  - The two-input switches are replaced by one-input switches which are reset by constant influxes.

```
directive sample 0.002 1000
directive plot x1(); y1(); b1(); x2(); y2(); b2(); c1();
c2()

val r = 10.0
new x1cat@r:chan
new y1cat@r:chan
new x2cat@r:chan
new y2cat@r:chan

val s = 10.0
new x1cat2@s:chan
new y1cat2@s:chan
new x2cat1@s:chan
new y2cat1@s:chan

val t = 100.0
new c1ch@t:chan new c2ch@t:chan

let c1gen() = delay@100000.0; (c1gen() | c1())
and c1() = !c1ch; ()
let c2gen() = delay@100000.0; (c2gen() | c2())
and c2() = !c2ch; ()

let x1() =
  do !x1cat; x1()
  or !x1cat2; x1()
  or ?c1ch; b1()
and y1() =
  do !y1cat2; y1()
  or ?x1cat; b1()
  or ?x2cat1; b1()
and b1() =
  do ?x1cat; x1()
  or ?x2cat1; x1()
  or ?c1ch; y1()

let x2() =
  do !x2cat; x2()
  or !x2cat1; x2()
  or ?c2ch; b2()
and y2() =
  do !y2cat1; y2()
  or ?x2cat; b2()
  or ?y1cat2; b2()
and b2() =
  do ?x2cat; x2()
  or ?y1cat2; x2()
  or ?c2ch; y2()

run 1000 of x1()
run 6666 of y1()
run 2333 of b1()

run 3333 of x2()
run 3333 of y2()
run 3333 of b2()

run 4000 of c1gen()
run 4000 of c2gen()
```





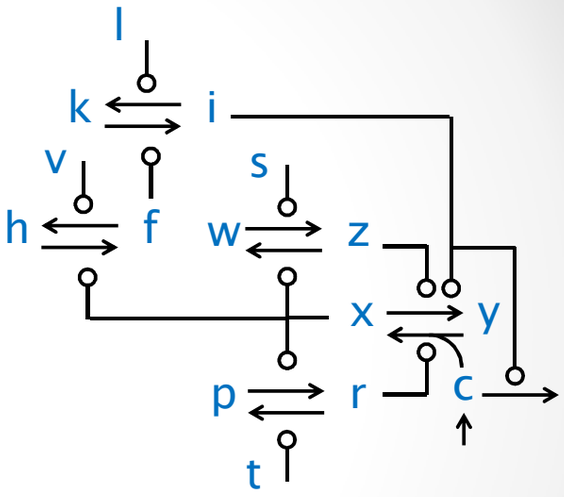


Works best with s=r.

Needs constant influx of c1,c2

Novak-Tyson Oscillator

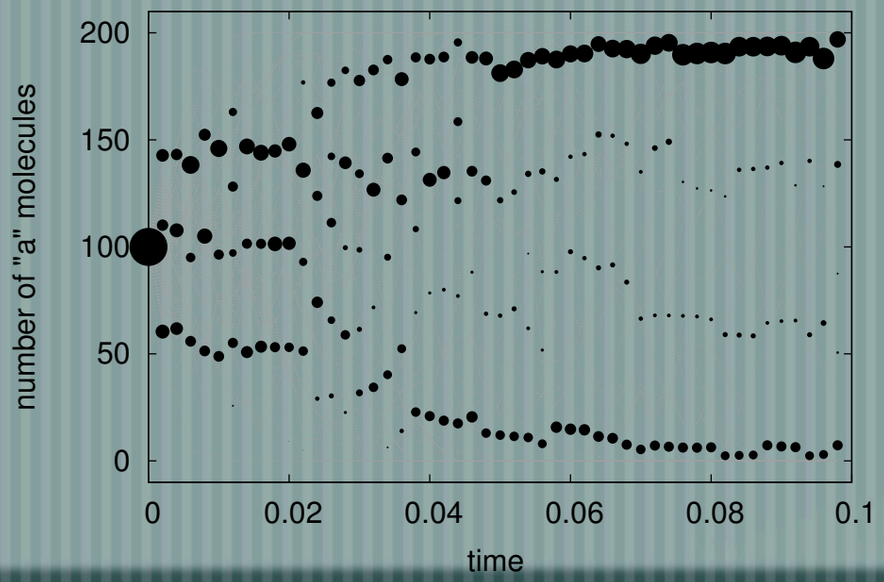
- First switch
  - Is the 'transformed' AM switch in one-input configuration (driven by constant influx of cyclin).
- Second switch
  - Is a simple two-stage switch working as a delay (the first switch is so good in terms of hysteresis that the second switch is not very critical for oscillation).
  - It can be replaced by a one-stage switch (Ferrell's cell cycle osciallor) but oscillation is a bit harder to obtain.
- Connection
  - Single links, as in the influx oscillator.

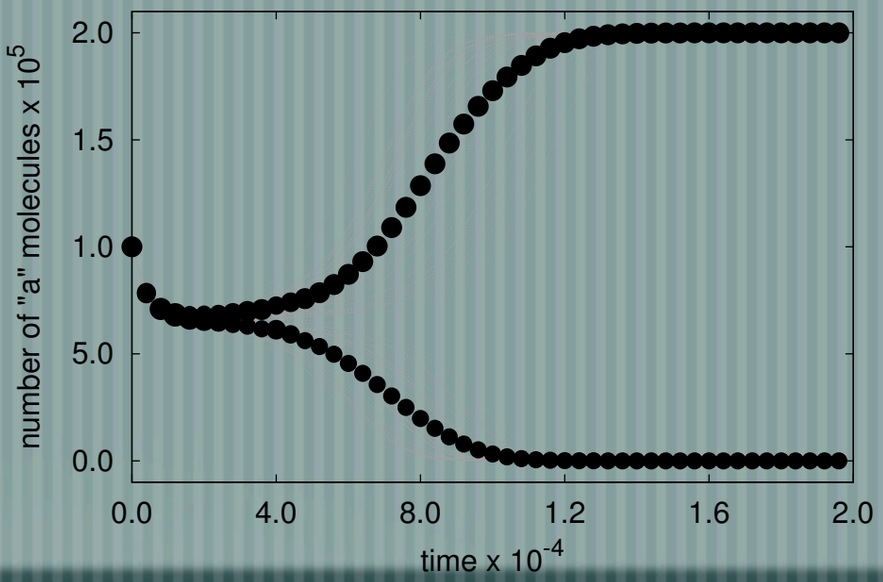


Journal of Cell Science 106, 1153-1168 (1993)
Printed in Great Britain © The Company of Biologists Limited 1993

Numerical analysis of a comprehensive model of M-phase control in *Xenopus* oocyte extracts and intact embryos

Bela Novak* and John J. Tyson†
Department of Biology, Virginia Polytechnic Institute and State University, Blacksburg, Virginia 24060-0406, USA



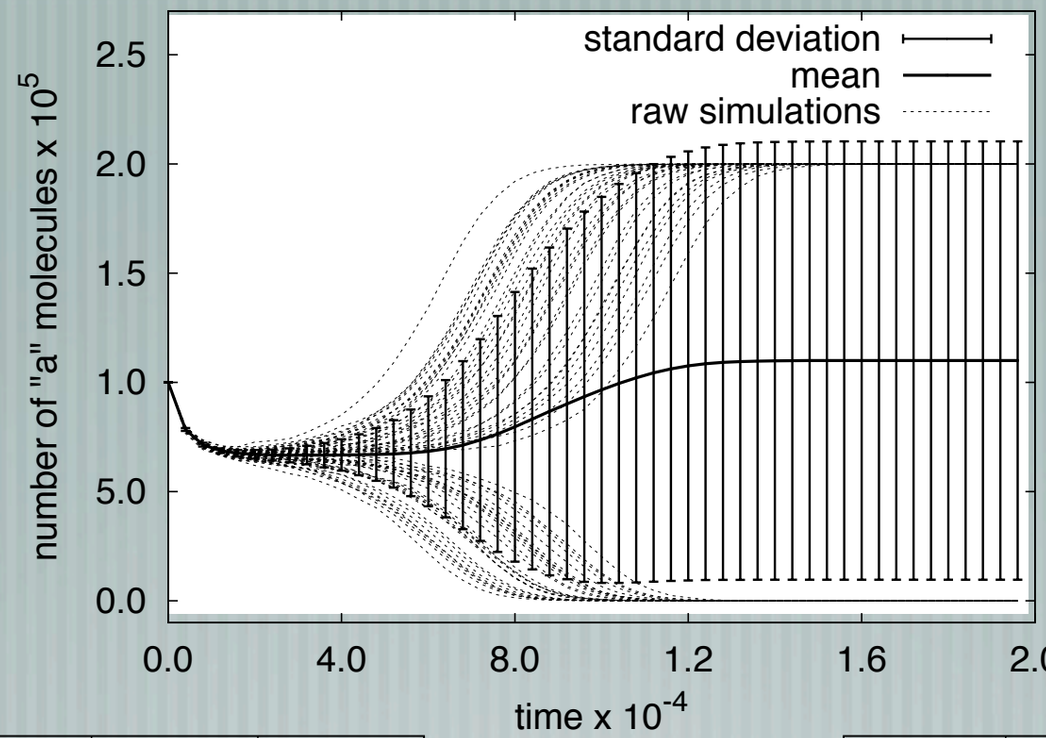


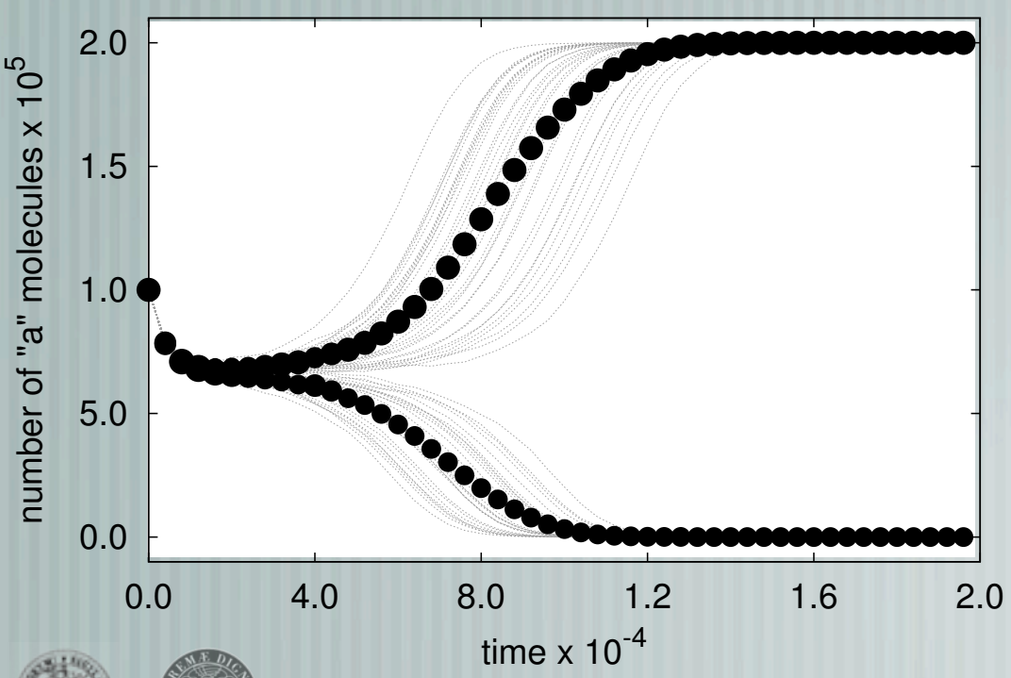
Demo

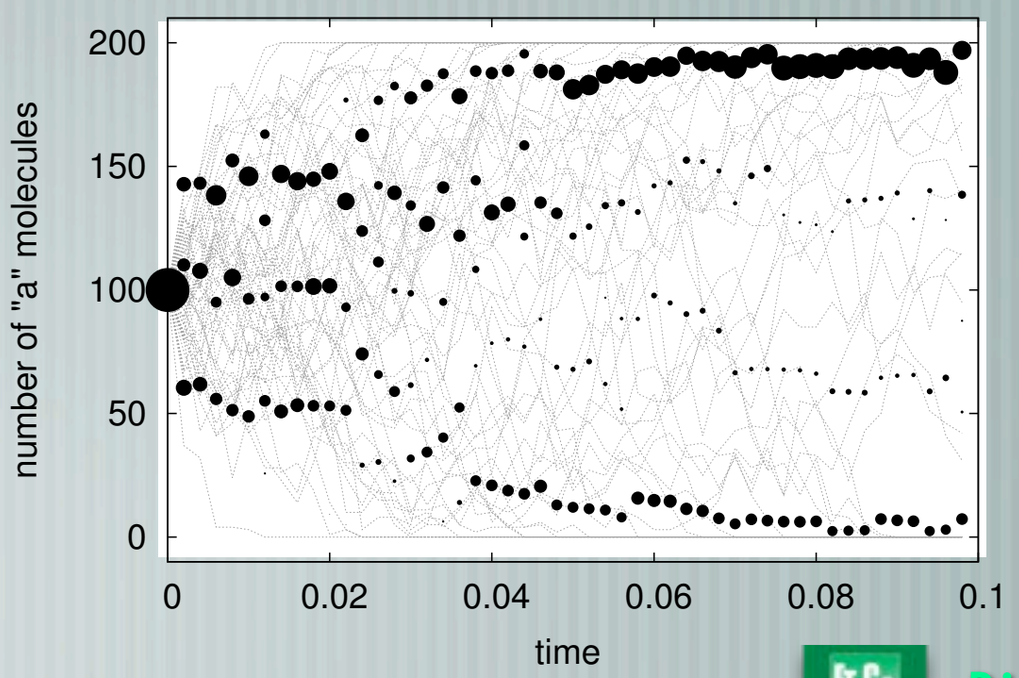
Graphical interface realised by
Etica SRL, Torino, Italy
http://www.eticasrl.com/

Tecnologie e servizi

Demo: unstable switch - stable switch

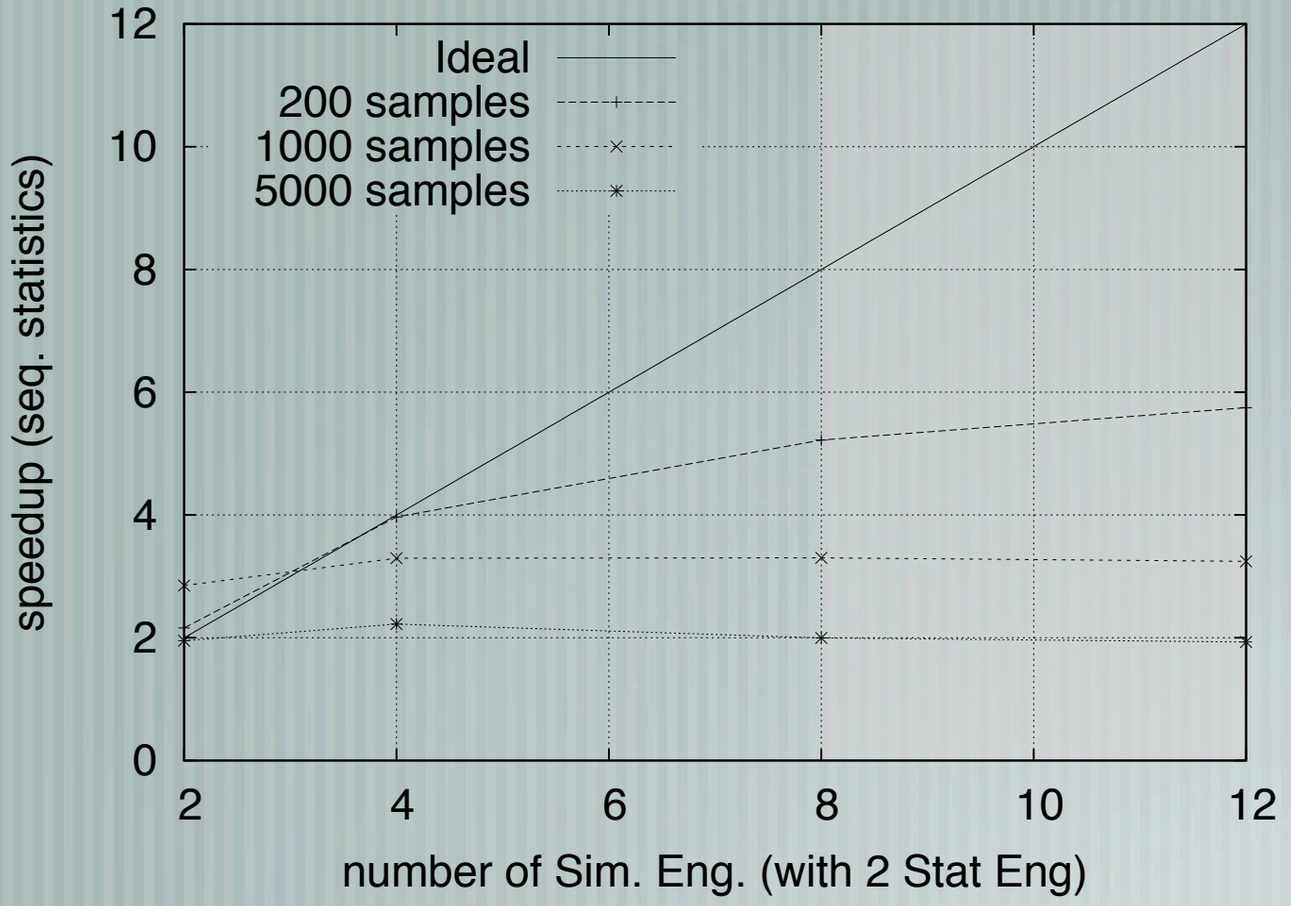


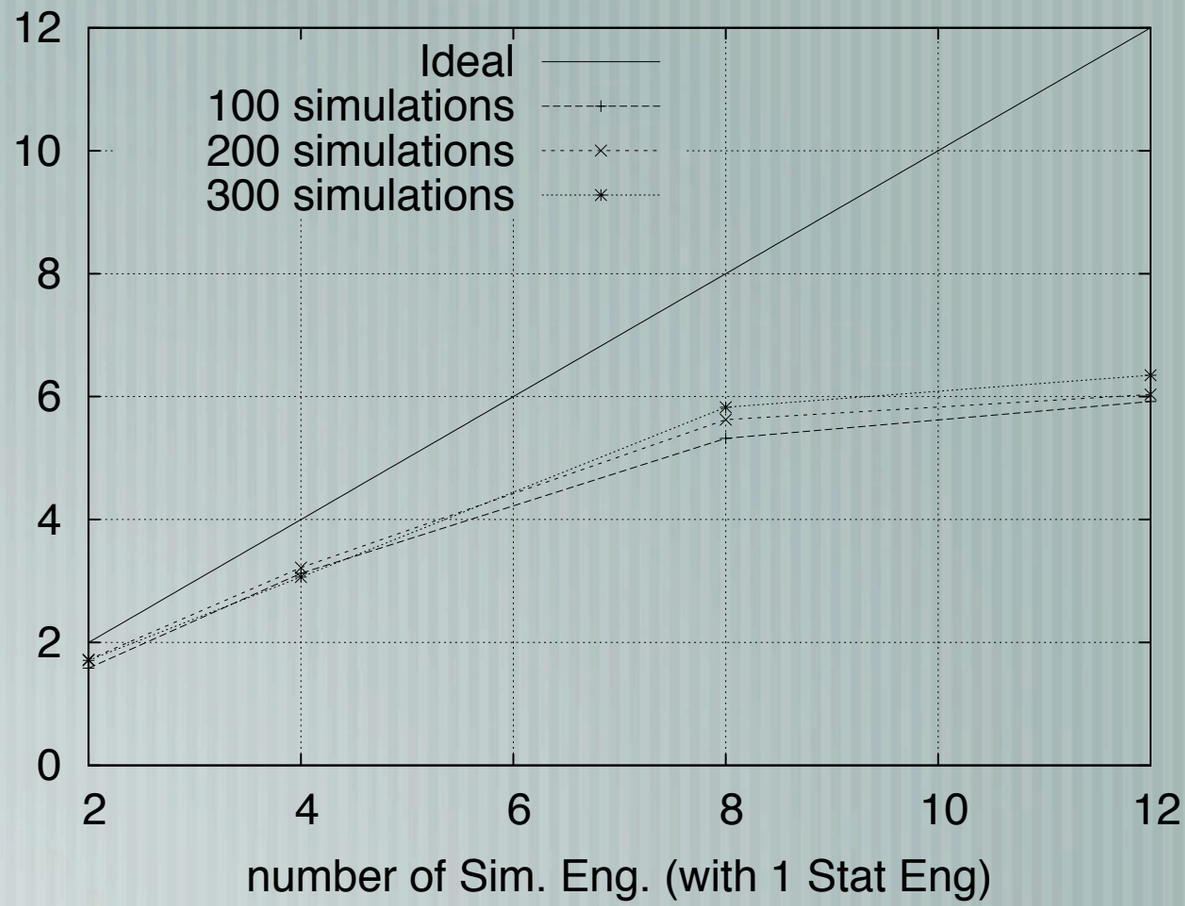




BioBITs

Speedup (Intel core 2x4-core)





BioBITs

Conclusions

- More details on FastFlow and software accelerators
  - Thursday Sept. 1st, 14.30 - Session E3 - room Denucé - Accelerating code on multi-cores with FastFlow.
  - HPC advisory council award at Intl. Supercomputing 2011

- Simulation and analysis tool
  - a design and developing methodology, not only a tool
    - easily adaptable to other Monte Carlo simulations (not only bio)
  - designed for multi-core, high-throughput, scalable, fast
  - open source and available for Mac OS, Linux, Windows
  - Graphical User Interface - allows to steer a remote a multi-core server




BioBITs