# Using queuing theory to model streaming applications

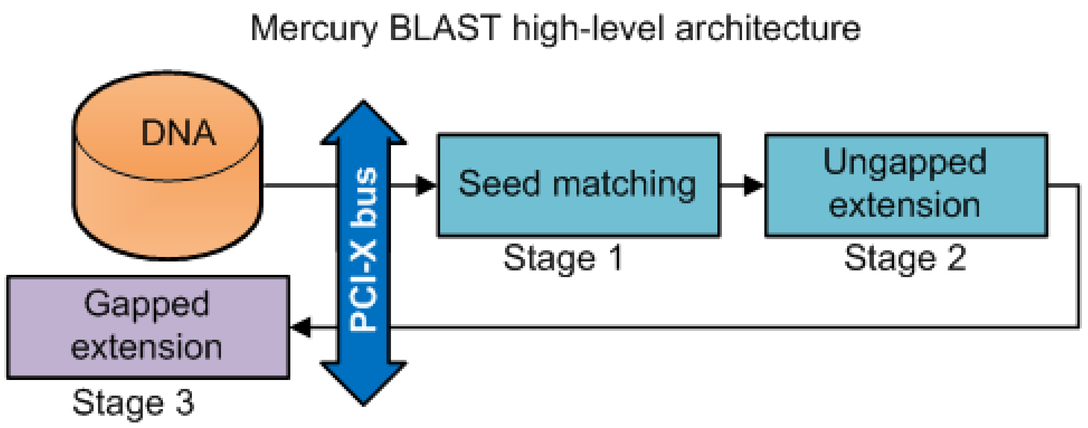
Rahav Dor, Joseph M. Lancaster, Mark A. Franklin, Jeremy Buhler, and Roger D. Chamberlain
Department of Computer Science & Engineering
Visit us at http://hpcb.wustl.edu

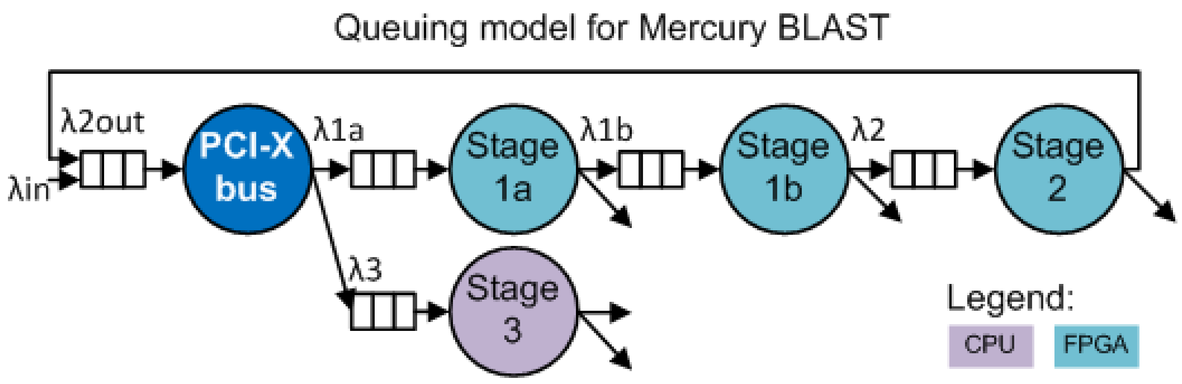
High Performance Computational Biology

SAAHPC July 2010

## 1 Research objective

- It is difficult to develop a mathematical model of an application
- Our goal is to define a modeling paradigm that is easy to develop and so truly appealing
- We would like to explore whether simple queuing models that do not necessarily match the system can be effective at describing the performance of real systems




## 2 Method

- Compare model to empirical results
- Construct a Jacksonian queuing network
  - Each stage is Markovian
  - Poisson arrival process with rate λ
  - Exponentially dist. service times with rate μ
  - Infinite queues
- We use TimeTrial (a minimal-impact performance measurement tool) to calibrate and validate the model
- We use two runs that represent distinct execution circumstances. Run 1 lightly taxes the system while run 2 heavily taxes at least stage 1b

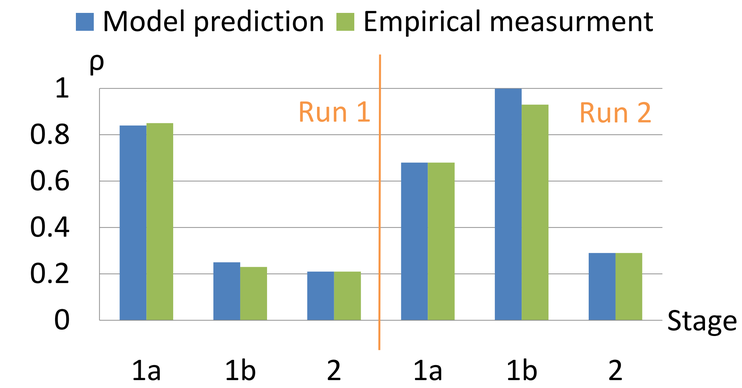
## 3 Model

- Queuing model M/M/1/∞
- Mercury BLAST properties
  - The actual application exhibits none of the model properties
  - Arrivals are not Poisson
  - Service time not exponential
  - Physical queues are finite

This work was supported by NSF and NIH

## 4 Results

**Stage utilizations, ρ**




- Stage utilizations are a close match in virtually every case
  - This validates input rates, service rates, and branching probabilities. But is not surprising as these directly determine ρ and are independent of the queuing model distributions

**Queue occupancy, $N_Q$**

| Queue | Run 1 | | Run 2 | |
|---|---|---|---|---|
| | Model prediction | Empirical result | Model prediction | Empirical result |
| $N_{Q,1b}$ | Near 0 | Near 0 | 7500 | 580 |
| $N_{Q,2}$ | Near 0 | Near 1 | Near 0 | Near 2 |

- Queue occupancies have very close match for three out of four cases
  - For $N_{Q,1b}$ in run 2 the high server utilization indicates this stage is the limiting bottleneck
  - The physical queue has 600 entries, so the empirical value cannot grow beyond that
  - The model predicts much higher occupancy, but corrects itself by predicting a high backpressure probability of $P_{BP,1b}[N>=600]=0.92$

## 5 Conclusions

- Seek more evidence for our hypothesis with additional experiments of Mercury BLAST and other applications
- These results illustrate that simple queuing models might be a good option for modeling streaming applications
- Where there are discrepancies, the model can assist in understanding them (e.g. $P_{BP,1b}$)