CONVOLVE

Dataflow Performance Modeling Tutorial

Marc Geilen, m.c.w.geilen@tue.nl
Electronic Systems, Dept. Electrical Engineering, Eindhoven university of Technology

Research funded by European Commission under grant agreement number 101070374

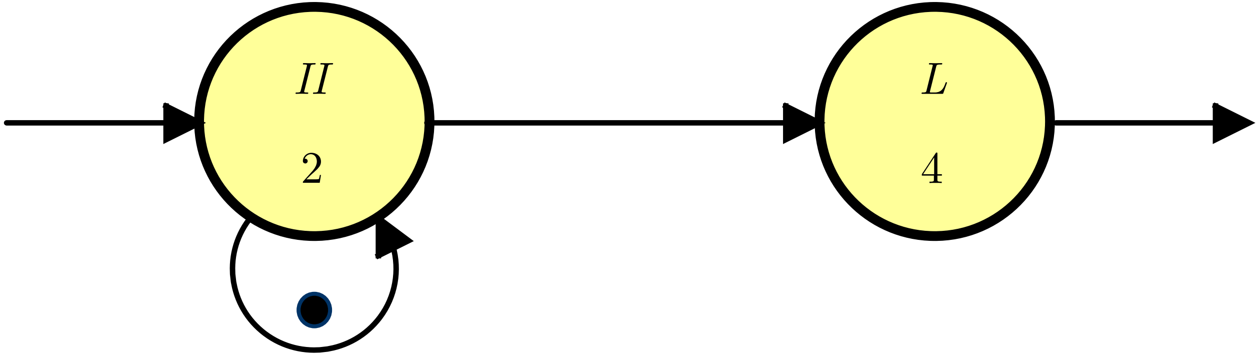
Context & Objectives

WP6, compositional analysis methods for design space exploration

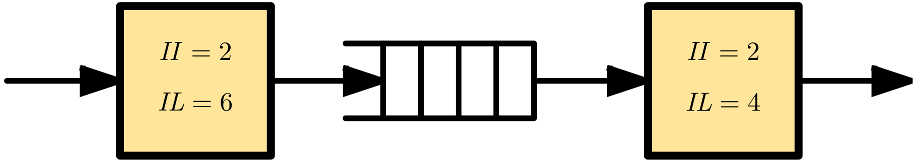
- Analysis techniques to assess performance of a proposed mapping
- Feedback about bottlenecks or critical paths
  - to support exploration

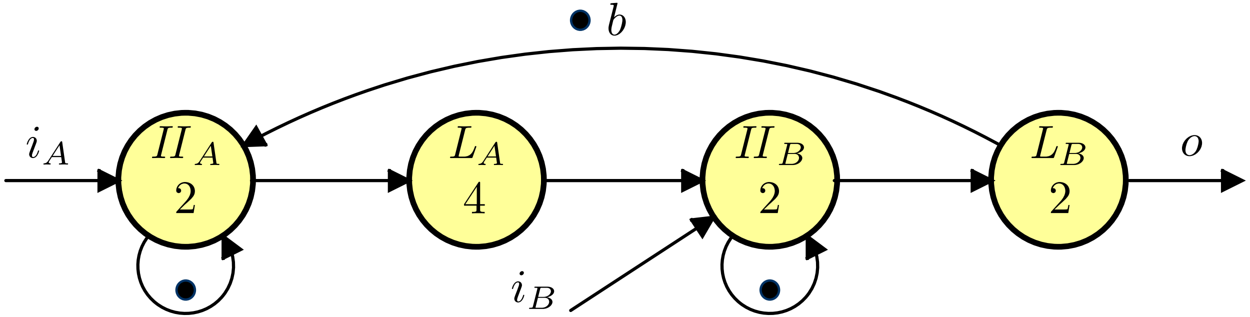
Dataflow Models

- **Dataflow**: model of activities and dependencies



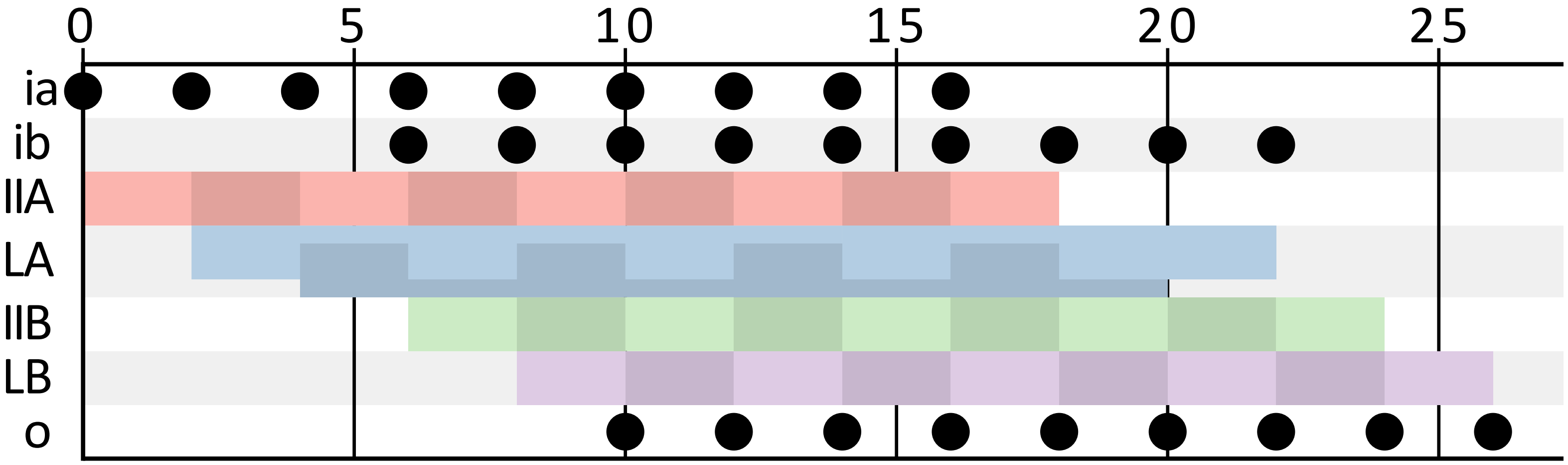
Example





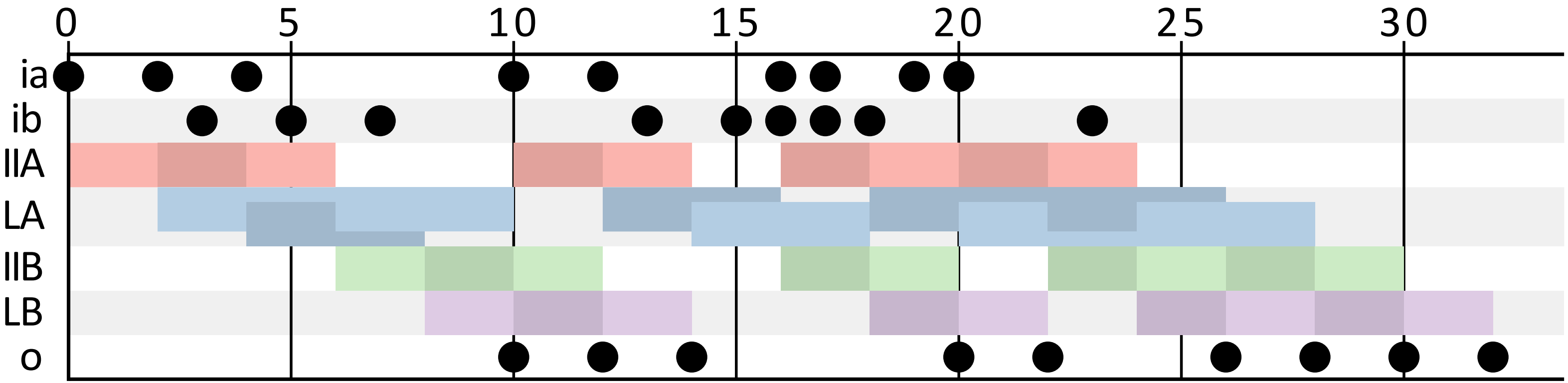
Gantt chart (1)

- Maximum throughput ASAP execution



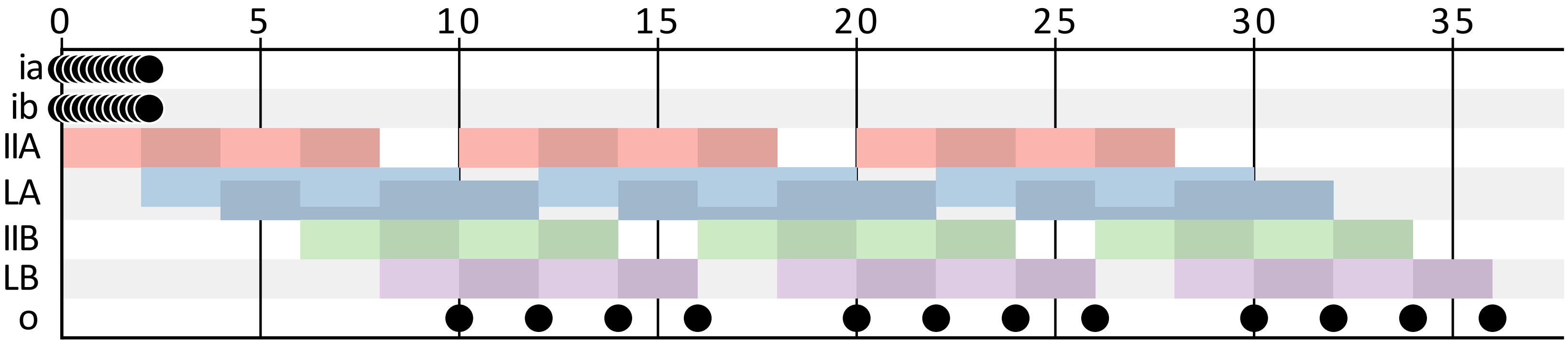
Gantt chart (2)

- Input dependencies



Gantt chart (3)

- Buffer capacity bottleneck



Max-plus Algebra

- a **linear algebra** for logistics
- $x \oplus y \otimes z = \max(x, y + z)$
- Including **matrix-vector calculus**
- **Linear system** with state matrix

$$A = \begin{pmatrix} 2 & -\infty & 2 & -\infty & -\infty & -\infty \\ 8 & 2 & 8 & -\infty & -\infty & -\infty \\ -\infty & -\infty & -\infty & 0 & -\infty & -\infty \\ -\infty & -\infty & -\infty & -\infty & 0 & -\infty \\ -\infty & -\infty & -\infty & -\infty & -\infty & 0 \\ 10 & 4 & 10 & -\infty & -\infty & -\infty \end{pmatrix}$$

Performance analysis

- **Throughput** is $\frac{1}{\lambda}$ if λ is the **largest eigenvalue** of the matrix
- **Latency** can be computed from state space matrices

$$\mathbf{\Lambda} = \mathbf{C}(-\mu \otimes \mathbf{A})^{*}\mathbf{B} \oplus \mathbf{D}$$

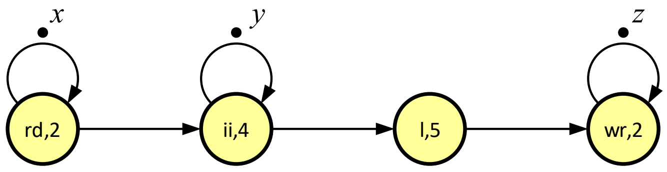
- Throughput (with buffer size 4) is $\frac{2}{5}$
- Latency for $i_A \rightarrow o$ is 10
- Latency for $i_B \rightarrow o$ is 4

Analysis

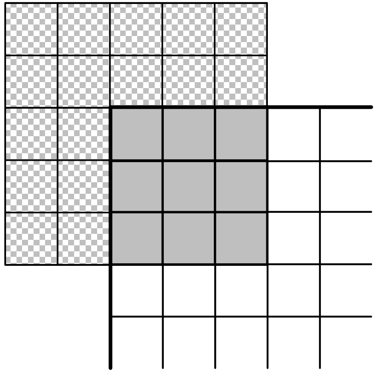
- Analysis provides performance numbers
- Models allow (automatic) exploration of **trade-offs** between **resource allocation** and **performance**
    - e.g., buffer size vs throughput
- feedback about performance bottleneck may provide guidance for design-space exploration.

Scaling and Dynamism

- we need to go to millions (?) of neurons
- multi-rate
- varying delays
- modes / scenarios

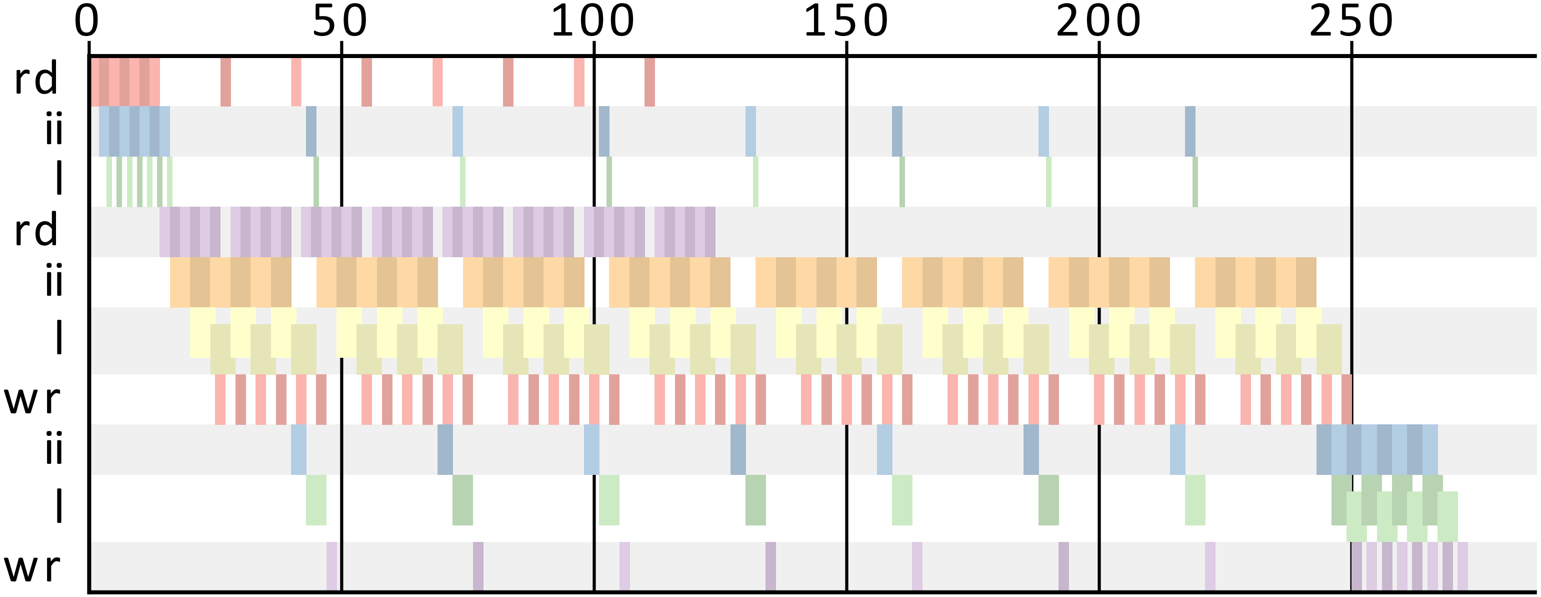


mode	rd	ii	l	wr
ri	2	2	0	0
cm	2	4	5	2
ro	0	3	4	2



Gantt chart

$$ri^6 \cdot (ri \cdot cm^6 \cdot ro)^8 \cdot ro^6$$



Compositionality

- Computing the overall max-plus matrix is still efficient

$$\mathbf{A}_{ri}^{6}\left(\mathbf{A}_{ri}\mathbf{A}_{cm}^{6}\mathbf{A}_{ro}\right)^{8}\mathbf{A}_{ro}^{6}$$

- Tracking critical path still possible
- Repetition patterns can be compositionally computed from modules

van der Vlugt, S., Alizadeh Ara, H., de Jong, R. et al. Modeling and Analysis of FPGA Accelerators for Real-Time Streaming Video Processing in the Healthcare Domain. J Sign Process Syst 91, 75–91 (2019). https://doi.org/10.1007/s11265-018-1414-3

Demo

- http://computationalmodeling.info/cmwb
- http://www.es.ele.tue.nl/sdf3
- https://github.com/Model-Based-Design-Lab/cmlib
- https://computationalmodeling.info/static/mpd/