Work-flows

Marco Danelutto
Dept. Computer Science
Univ. of Pisa

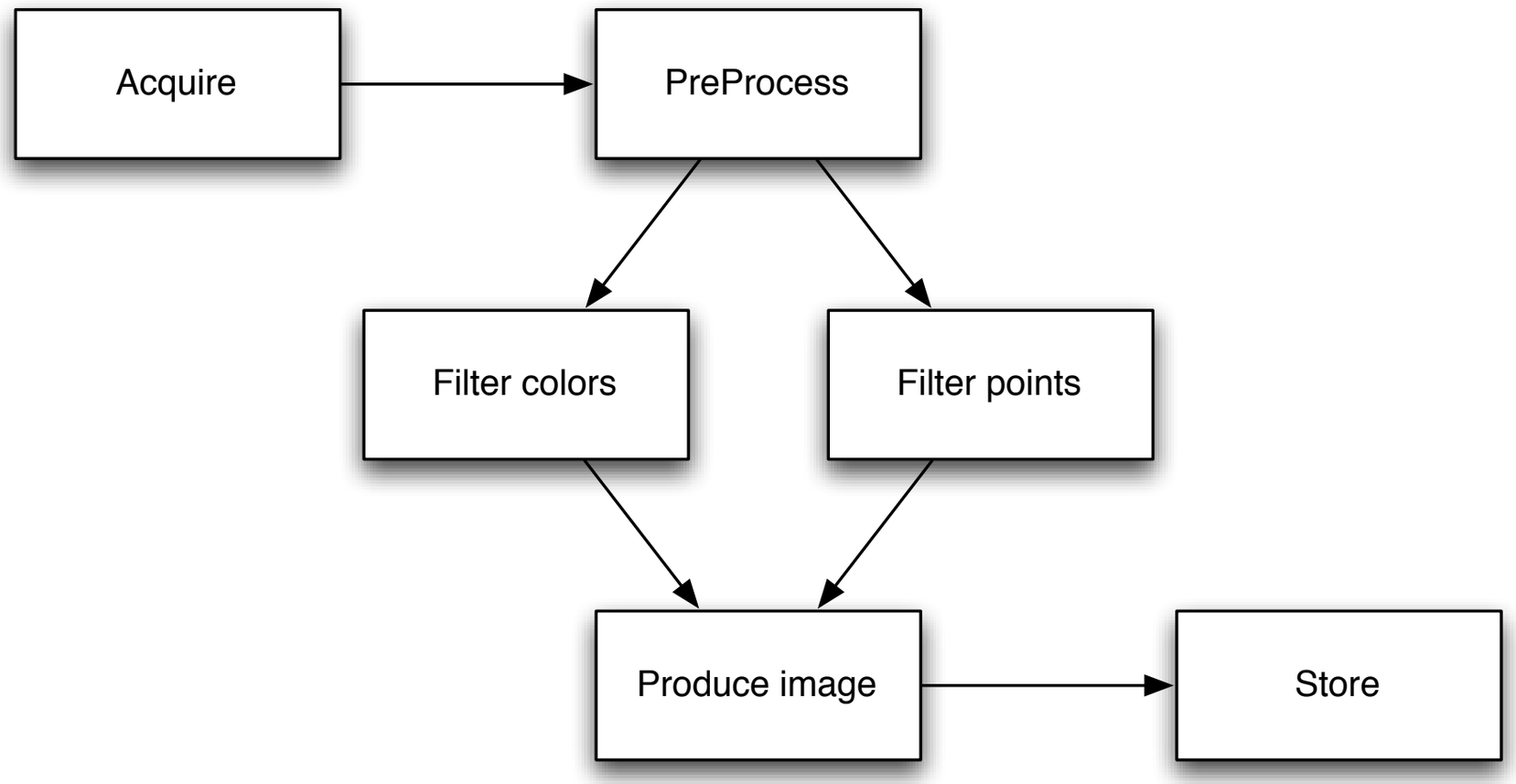
Master Degree (Laurea Magistrale) in
Computer Science and Networking
Academic Year 2009-2010

Concept of workflow

- **Sequence of steps (usually called TASKS)**
- **With data and control flow connecting these steps**
- **Usually**
  - *each step can be computationally intensive*
  - *amount of data passed between stages could be huge*
- **R. Buyya "*A Taxonomy of Scientific Workflow Systems for Grid Computing*" 2005**
  - *Scientific workflow is concerned with the automation of scientific processes in which tasks are structured based on their control and data dependencies.*

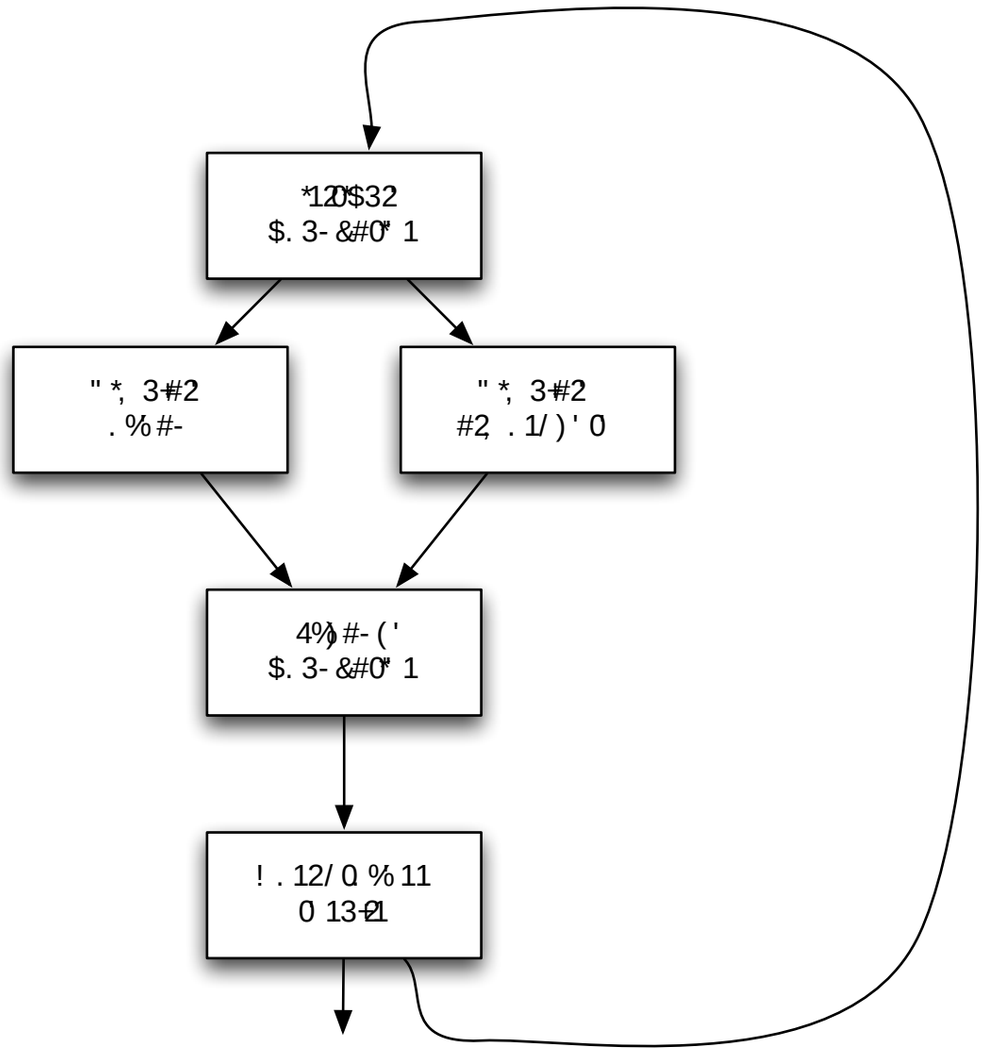
Sample Workflow

- **Arcs represent in this case both data and control dependencies**



Workflow structure

- **DAG or non DAG (Direct Acyclic Graphs)**
  - *DAG: directed graph with no directed cycles (no way to come back to the same node following arcs)*
- **most problems represented as DAGs**
- **several iterative problems are not DAGs**



Workflow model/specification

- **Abstract**
  - *DAG or non-DAG graph*
  - *expresses logical dependencies*
- **Concrete**
  - *adds to abstract*
    - tasks to resource binding
    - communication & synchronization (control) details
- **Concrete model**
  - *usually compiled from the abstract one*
  - *when resources are known*

Workflow scheduling

- **mapping and managing execution of workflow tasks**
- **on available resources**
- **alternatives**
  - *centralized (better decisions, not scalable)*
  - *decentralized (worse decisions, scalable)*
  - *hierarchical (acceptable decisions, acceptable scalability)*
  - ***in principle !!!*** *(not true for small workflows)*

Workflow scheduling (2)

- **Static decisions**
  - *based on the workflow graph*
  - *not influenced by resource dynamicity*
  - *nor by past history of workflow execution*
- **Dynamic decisions**
  - *re-using history from the past executions*
  - *using simulation results*
    (based on predicted task execution times)
- **Just in time scheduling**
  - *probably “locally (in time) better” resource usage*

Scheduling strategies

- **Performance driven**
  - *pick up better resources for computationally intensive tasks*
- **Market driven**
  - *use market models to schedule tasks to resources*
    - most powerful resources have a high price => market driven scheduling may produce “slower” results at a smaller price
- **Trust driven**
  - *when security is crucial*
  - *do not select untrusted host or host with past “failures”*