

Fluxo de Dados em Workflows Científicos: O Sistema KEPLER

Aluno: William Tankiti Yamamoto (wt@linux.ime.usp.br)

Orientador: João Eduardo Ferreira (jef@ime.usp.br)

Departamento de Ciência da Computação
Instituto de Matemática e Estatística – Universidade de São Paulo

Definição de Workflow

É a automação do processo de negócio onde documentos, informações ou tarefas são passadas de um participante para o outro de acordo com um conjunto de regras de procedimentos para se atingir um objetivo de negócio.

O que é um Workflow Científico?

É a aplicação do conceito de *workflow* em ambientes científicos. Diferentemente dos ambientes de negócio, os ambientes científicos visam a análise e a manipulação de grande volume de **dados**.

Workflow Científico vs. Workflow de Negócio

Ambiente Científico



Controle de fluxo de dados (DATAFLOW)

- Orientado a dado
- Grande volume de dados

Ambiente de Negócios



Controle de fluxo de atividades (CONTROLFLOW)

- Orientado à atividade
- Pequena escala de atividades

Motivação

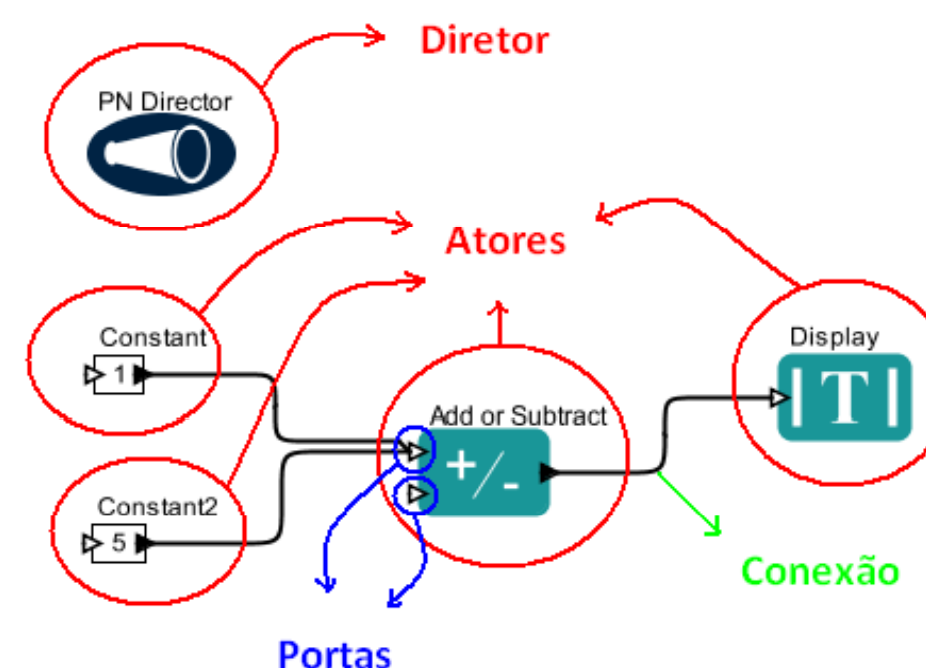
Combinação do controle de fluxo de dados com o fluxo de atividades. **Entender o funcionamento do fluxo de dados no workflow científico é imprescindível para a integração dos dois ambientes.**

Sobre o KEPLER

É um ambiente de desenvolvimento específico para análise, manipulação, modelagem e simulação de *workflows científicos*. Atualmente, o KEPLER é utilizado em áreas como a bioinformática, ecoinformática e geoinformática, proporcionando aos cientistas uma ferramenta de criação e execução de seus experimentos. O KEPLER foi desenvolvido com base nos **princípios da orientação a atores**.

Modelo Orientado a Atores

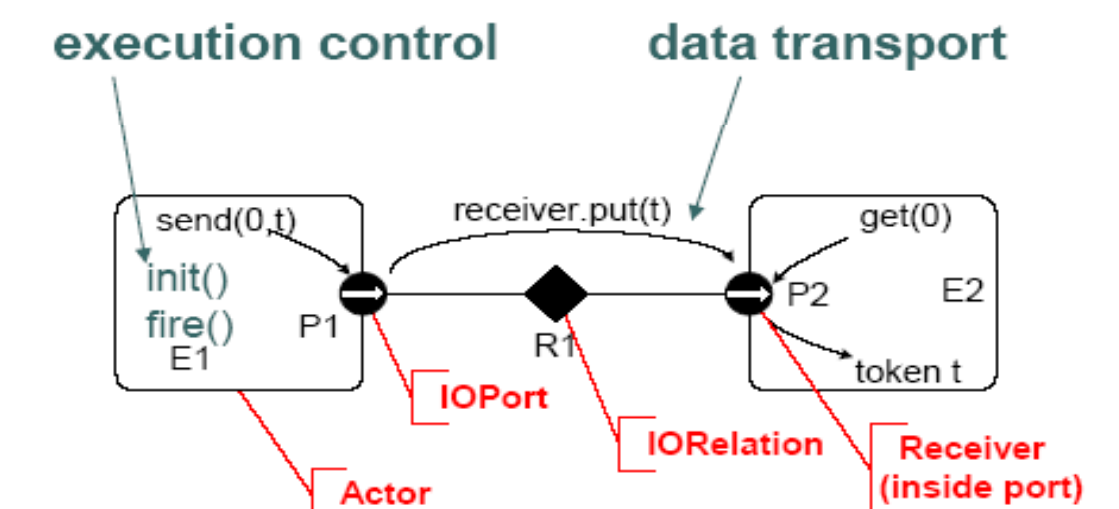
Sustentado pelo modelo inicialmente proposto por Hewitt [4] e posteriormente estendido e formalizado por Agha [5], KEPLER utiliza a metáfora diretor/ator para representar os componentes do workflow. Assim como em um filme, o diretor do KEPLER comanda o seu elenco (atores), especificando **quando** cada um deve agir e **como** estão conectados entre si. Já os atores **processam** os dados disponíveis nas suas portas de entrada e disponibilizam os resultados na porta de saída.



Componentes de um workflow representados no KEPLER

Fluxo de Dados no KEPLER

O fluxo de dados (*dataflow*) pode ser representado pela figura abaixo:



* Desenho retirado de
<http://embedded.eecs.berkeley.edu/concurrency/lectures/Lee05.pdf>

O ator origem E1 tem uma porta de saída P1, indicada na figura com uma flecha na direção do fluxo do **token**. O ator destino E2 tem uma porta de entrada P2.

E1 chama o método **send()** de P1 para enviar um token *t*. A porta obtém a referência ao receptor remoto (através da relação R1) e chama o método **put()** do receptor, passando-lhe o token. O ator destino E2 recupera o token *t* chamando o método **get()** de sua porta de entrada, a qual, por sua vez, chama o método **get()** do receptor designado.

Referências Bibliográficas

- [1] S. Bowers and B. Ludaescher. Actor-oriented design of scientific workflows. In *International Conference on Conceptual Modeling (ER'2005)*, 2005
 - [2] Edward A. Lee and Stephen Neuendorffer, "Concurrent Models of Computation for Embedded Software", *Technical Memorandum UCB/ERL M04/26*, University of California, Berkeley, CA 94720, July 22, 2004.
 - [3] Grace A. P. Borges, "Controle de dados em Processos de Negócio e Workflows Científicos", *Exame de Qualificação de Mestrado*, 2007
 - [4] Carl Hewitt. *et al.* Actor Induction and Meta-evaluation, Conference Record of ACM Symposium on Principles of Programming Languages, January 1974
 - [5] Gul Agha. Actors: A Model of Concurrent Computation in Distributed Systems, Doctoral Dissertation, 1986
 - [6] Workflow Management Coalition, 1995
- Mais informações sobre o Kepler <http://www.kepler-project.org>