scICN: Scientific Information Control Nets

Hyun Ahn*, Minjae Park**, Kwanghoon Pio Kim*

*Dept. of Computer Science, KYONGGI UNIVERSITY, Suwon Kyonggido, Korea
**BISTel, Inc., Seoul, Korea
{hahn, kwang}@kgu.ac.kr, mjpark@bistel-inc.com

***Abstract*—In this paper, we propose a formal description for representing the scientific workflow model supporting process-driven scientific behaviors in data intensive experiment procedures and large scale computing environments. We try to make a conceptual extension of the information control net methodology so as to be applicable to the scientific workflow models and systems, and dub it scICN (Scientific Information Control Net). Upon the scICN-based scientific workflow model, we exemplify its application to a pseudo scICN-based scientific workflow model.**

***Keywords*-scientific workflow; dominancy; load-balancing; verification; exception handling and recovery**

I. Introduction

Traditionally, the goal of a workflow management system was to describe, control, and monitor the enactments of business procedures in a workflow-supported organization. According to flourish on the almost all industries through the various types of process-driven business models, such as process portals, POD (Process-on-Demand), process choreography and orchestration, process collaboration, and inter-organizational workflows, the applicability of the workflow management system has been changing, broadening, and scaling up, too, along with dramatic progressions of workflow and business process technologies. In recent, what we see happening now in the workflow literature is that the applicability reaches up to the scientific knowledge discovery arena, which is so-called "Scientific Workflow[1][2]," that is able to not only support the whole stage of scientific experiments and simulations but also automate exploratory processes for discovering scientific knowledge involving cycles of observation, hypothesis formation, experiment design and execution. At this point, we would faithfully follow the conceptual definition of the scientific workflow stated in [1],[2], and [3], the like of which the scientific workflow implies the large scale and data intensive workflow, in terms of describing and deploying the overall idea and the basic concepts throughout the paper.

In this paper, we particularly focus on the quality of service issues in scientific workflow technologies, in terms of efficient load-balancing strategies in designing a large scale scientific workflow enactment architecture, and effective exception handling and recovery strategies in implementing a data intensive scientific workflow system. In order to cope with these quality of service issues, we need a well-defined scientific workflow meta-model that is surely applicable to the scientific workflow model as well as the architectural functionality of the underlying scientific workflow system. As we know, there exist tens of proposals[4][5][6][7] and rationales[3][8] about defining the concept of scientific workflow models and systems, so far. However, we, in this paper, try to newly define a scientific workflow meta-model rather than choosing one of those existing scientific workflow models. We also extend the conventional information control net methodology[9] to graphically and formally represent the scientific workflow model spawned from the defined meta-model, which is dubbed "Scientific Information Control Net" that is abbreviated to **scICN**, from now on.

In organizing the paper, we start from defining the basic concept of the scientific workflow meta-model with a series of functional definitions of the scientific activity types. The next section formalizes the scientific information control net and applies its graphical and formal representation to a pseudo scientific workflow model. With describing the implications of the proposed mathematical formalism for large scale and data intensive scientific workflows, we finalize the paper.

II. Scientific Workflow Model

In this section, we define a scientific workflow meta-model for modeling scientific experimental processes characterized by the large scale and data intensive properties. Based upon the meta-model, we are able to instantiate scICN-based scientific workflow models that are conceptually extended from the information control net methodology[9][10].

A. Meta-Model

In order to clearly define a scientific workflow model, it is needed to design a meta-model that is used to be signified by a set of entity types and their relationships. Furthermore, we have to take account of the conceptual definition of scientific workflows to conceive a reasonable and applicable meta-model. In general, the scientific workflow model is a mathematical methodology to specify various phases of the large and complex science process involving cycles of modeling and automation of computational experiments, data analysis, and data management. The well-described conceptual definition of the scientific workflows was introduced in the project DAKS[3] as followings:

"*A scientific workflow is the description of a process for accomplishing a scientific objective, usually expressed in terms of tasks and their dependencies. Typically, scientific workflow tasks are computational steps for scientific simulations or data*

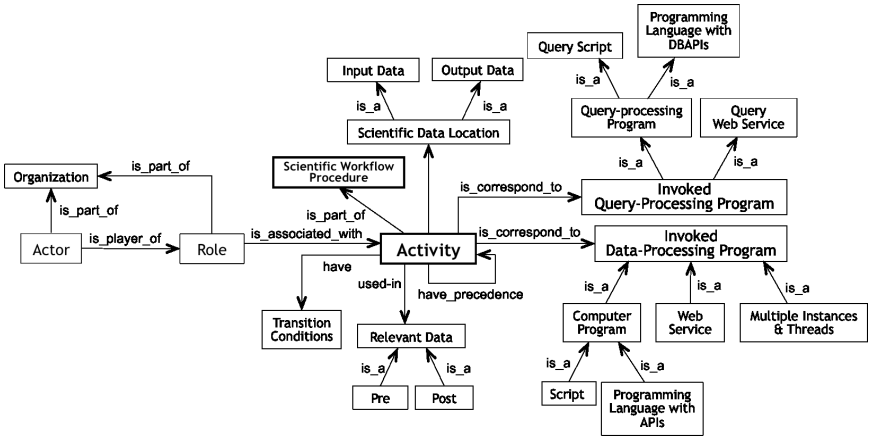


Fig. 1. Meta-model for Modeling Scientific Workflows

analysis steps. Common elements or stages in scientific workflows are acquisition, integration, reduction, visualization, and publication (e.g., in a shared database) of scientific data. The tasks of a scientific workflow are organized (at design time) and orchestrated (at runtime) according to data-flow and possibly other dependencies as specified by the workflow designer. Workflows can be designed visually, e.g., using block diagrams, or textually using a domain-specific language."

We design a scientific workflow meta-model by faithfully reflecting the above definition as much as possible. Figure 1 depicts the scientific workflow meta-model drawn from in-depth discussions in the authors' collaborative research groups. The primitive entity types constituting the scientific workflow model are Activity and Tasks (Invoked Query-processing program and Invoked Data-processing program) entity types; the primitive entity types has also certain associations with their own supplementary entity types, like Role, Data Location, and Relevant Data. The followings are the basic definitions of the primitive entity types:

- An **Activity** is a conceptual entity of the basic unit of work (computational task or step), and the activities in a scientific workflow model have precedence relationships, each other, according to their execution sequences based upon data-flow dependencies and/or a temporal ordering of computational steps. Also, the activity entity can be precisely specified by one of the elementary entity types, G-type, D-type, Q-type, and DQ-type, as depicted in Figure 2.
  - The elementary activity implies a computational step that can be realized by either data-processing (D-type), query-processing (Q-type), or both (DQ-type) applications implemented through computer programs, threads, transactions, query-scripts, or web services. Especially, we suppose that the D-type and Q-type activities can be realized by applying the concepts of multiple instance patterns[3], like Multiple Instances with a Priori Design Time Knowledge and Multiple Instances with a Priori Run Time Knowledge implemented in Kepler[5][11]. Also, the D-type data processing activity implies a kind of computational steps for scientific simulations or data analysis tasks performing acquisition, integration, reduction, visualization, or publication (e.g., in a shared database) of scientific data.
  - Particularly, the gateway (G-type) activity means a pair of split-join pseudo step that is used to controlling execution sequences of elementary/compound computational steps. Also, we assume that each of the gateway steps forms in the structured properties, *matched pair* and *proper nesting*. There are three sorts of G-type activities, such as conjunctive gateway (paralleling), disjunctive gateway (alternating), and iterative (looping) gateway. Particularly, the disjunctive gateway needs to be set some specific **Transition Conditions** in order to select one of the possible transition paths during the execution time. The transition condition itself can be defined by using certain input/output relevant data registered on the **Relevant Data** repository.
  - Additionally, there may be a special computational step, the compound activity, that signifies a computational step containing another scientific workflow model, which is so-called **Subprocess**.

- A **Scientific Workflow Procedure** is defined by a predefined or intended set of tasks or computational steps, called Activities, which reflects the activities' temporal ordering of executions and/or data-flow dependencies. A system, so-called scientific workflow management system, helps to organize, control, execute,and monitor such defined scientific workflow models. Conclusively, a scientific workflow model can be described by a temporal order of the associated computational steps through the combinations of sequential logics, conjunctive logics (after activity A, do activities B and C, simultaneously), disjunctive logics (after activity A, do either activity B or C, alternatively), and iterative logics, as illustrated in Figure 3.

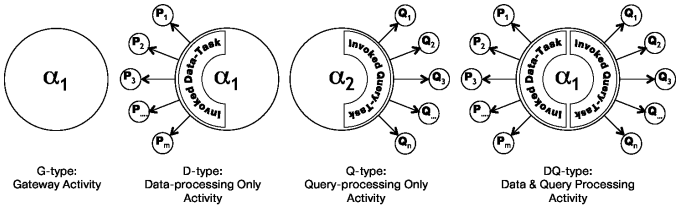


Fig. 2. Activity Types of Scientific Workflows

B. Scientific Information Control Net

In this paper, we would extensively adopt the original information control net (ICN) methodology[9] as the scientific workflow modeling methodology, which is dubbed to scICN-based scientific workflow model upon the meta-model described in the previous subsection. Based upon the scientific information control net, we are going to dig up a possible way of analyzing activity dominancies on a scientific workflow model in the next section. The scientific workflow model needs to be described by a formal representation so that it is able to provide a means to eventually specify the model either in textual language or in database format, or in both. The following definition is the formal representation of the scientific information control net:

Definition 1: scICN: **Scientific Information Control Net** of the scientific workflow model. A basic scICN is 10-tuple $\Gamma = (\delta, \gamma, \lambda, \theta, \varepsilon, \pi, \vartheta, \kappa, \mathbf{I}, \mathbf{O})$ over a set **A** of activities (including a set of compound/elementary/gateway computational steps), a set **T** of transition conditions, a set **R** of relevant data, a set **D** of invoked data-task applications, a set **Q** of invoked query-task applications, a set **P** of scientific roles, a set **L** of scientific data locations, and a set **C** of participants, where $\wp()$ is the Power-set function.

- **I** is a finite set of initial input data locations, assumed to be loaded with information by some external scICNs before execution of the corresponding scICN;
- **O** is a finite set of final output data locations, perhaps containing information used by some external scICNs after execution of the corresponding scICN;
- $\delta = \delta_i \cup \delta_o$
  where, $\delta_o : \mathbf{A} \longrightarrow \wp(\mathbf{A})$ is a multi-valued mapping function of a computational step to its set of (immediate) successors,
  and $\delta_i : \mathbf{A} \longrightarrow \wp(\mathbf{A})$ is a multi-valued mapping function of a computational step to its set of (immediate) predecessors;
- $\gamma = \gamma_i \cup \gamma_o$
  where $\gamma_o : \mathbf{L} \longrightarrow \wp(\mathbf{A})$ is a multi-valued mapping function of a computational step to its set of output data locations,
  and $\gamma_i : \mathbf{L} \longrightarrow \wp(\mathbf{A})$ is a multi-valued mapping function of a computational step to its set of input data locations;
- $\lambda = \lambda_a \cup \lambda_p$
  where $\lambda_p : \mathbf{D} \longrightarrow \wp(\mathbf{A})$ is a single-valued mapping function of a computational step to its invoked data-task application with multiple threads,
  and $\lambda_a : \mathbf{A} \longrightarrow \wp(\mathbf{D})$ is a multi-valued mapping function of an invoked data-task application with multiple threads to its set of associated computational steps;
- $\theta = \theta_a \cup \theta_p$
  where $\theta_p : \mathbf{Q} \longrightarrow \wp(\mathbf{A})$ is a single-valued mapping function of a computational step to its invoked query-task application with multiple threads,
  and $\theta_a : \mathbf{A} \longrightarrow \wp(\mathbf{Q})$ is a multi-valued mapping function of an invoked query-task application with multiple threads to its set of associated computational steps;
- $\varepsilon = \varepsilon_a \cup \varepsilon_p$
  where $\varepsilon_p : \mathbf{P} \longrightarrow \wp(\mathbf{A})$ is a single-valued mapping function of a computational step to one of the scientific roles,
  and $\varepsilon_a : \mathbf{A} \longrightarrow \wp(\mathbf{P})$ is a multi-valued mapping function of a scientific role to its sets of associated computational steps;
- $\pi = \pi_p \cup \pi_c$
  where, $\pi_c : \mathbf{C} \longrightarrow \wp(\mathbf{P})$ is a multi-valued mapping function of a scientific role to its set of associated scientists,
  and $\pi_p : \mathbf{P} \longrightarrow \wp(\mathbf{C})$ is a multi-valued mapping function of a scientist to its sets of associated scientific roles;
- $\vartheta = \vartheta_i \cup \vartheta_o$
  where $\vartheta_o : \mathbf{R} \longrightarrow \wp(\mathbf{A})$ is a multi-valued mapping function of a computational step to its set of output relevant data,
  and $\vartheta_i : \mathbf{R} \longrightarrow \wp(\mathbf{A})$ is a multi-valued mapping function of a computational step to its set of input relevant data;
- $\kappa = \kappa_i \cup \kappa_o$
  where $\kappa_i(\alpha)$: sets of control-transition conditions, **T**, on each arc, $(\delta_i(\alpha), \alpha), \alpha \in \mathbf{A}$;
  and $\kappa_o(\alpha)$: sets of control-transition conditions, **T**, on each arc, $(\alpha, \delta_o(\alpha)), \alpha \in \mathbf{A}$;
  where the set $\mathbf{T} = \{default, or(conditions), and(conditions)\}$.

In terms of the graphical representation, we principally adopt the original notations of information control net[9][10]. As depicted in Figure 3, there are possible primitive computation-flows and their related gateway G-type activities. That is, the conjunctive (or parallel) computation-flow type with a pair of conjunctive gateway G-type activities, split-AND and join-AND, is represented by solid dots(•), meanwhile the disjunctive (or decision) computation-flow type with a pair of disjunctive gateway G-type activities, split-OR and join-OR, is represented by hollow dots(○). Also, the iterative (loop) computation-flow type with a pair of loop gateway G-type activities, split-LOOP and join-LOOP, is represented by double hollow dots.

Besides, in order to be syntactically safe, it is very important for these gateway G-type activities to keep the structured prop-

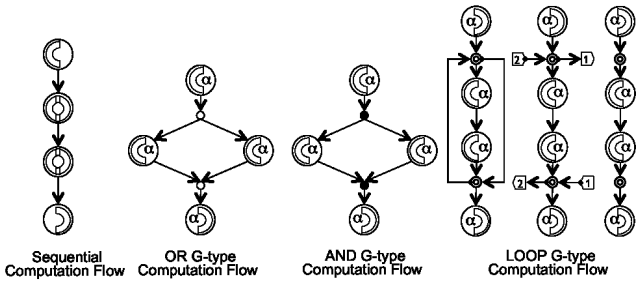


Fig. 3. Logics of Temporal Orders with Gateway Activities in Scientific Workflows

erties—proper nesting and matched pair properties. Therefore, in specifying a scientific information control net, not only each of the gateway G-type activities always keeps matched pair with split and join types, but also multiple sets of the gateway G-type activities keep in a properly nested pattern.

III. Conclusions

So far, this paper has expatiated the formal scientific information control net and its related definitions by extending the original information control net methodology. Based upon these formal and graphical definitions, we are able to construct a scientific workflow model. In recent, the literature needs various, advanced, and specialized scientific workflow modeling methodologies that are well-suitable to the domain of data intensive and very large scale scientific computing environments. We so strongly believe that this work might be one of those impeccable attempts and pioneering contributions for improving and advancing the large scale and data intensive and scientific workflow management technology.

Acknowledgements

This research was supported by the Basic Science Research Program through the National Research Foundation of Korea(NRF) funded by the Ministry of Education, Science and Technology (Grant No. 2012006971), Republic of Korea.

References

[1] A. Barker and J. van Hemert, "Scientific workflow: A survey and research directions," *Lecture Note in Computer Science*, vol. 4967, pp. 746–753, 2008.

[2] J. Qin and T. Fahringer, *SCIENTIFIC WORKFLOWS: Programming, Optimization, and Synthesis with ASKALON and AWDL*. Springer, 2012.

[3] A. G. Ustun Yildiz and A. H. Ngu, "Business versus scientific workflow: A comparative study," *University of California Davis, Department of Computer Science, Research Report, Project DAKS*, vol. n 2009-3, pp. 1–18, 2009.

[4] N. H. T. M. Anne H.H. Ngu, Shawn Bowers and T. Critchlow, "Flexible scientific workflow modeling using frames, templates, and dynamic embedding," *Lecture Note in Computer Science*, vol. 5069, pp. 566–572, 2008.

[5] e. a. Bertram Ludascher, "Scientific workflow management and the kepler system," *Journal of Concurrency and Computation: Practice & Experience*, vol. 18, no. 10, pp. 1039–1065, August 2006.

[6] Z. Y. Ping Yang and S. Lu, "Formal modeling and analysis of scientific workflows using hierarchical state machines," in *Proceedings of IEEE International Conference on e-Science and Grid Computing*. IEEE, December 2007, pp. 619–626.

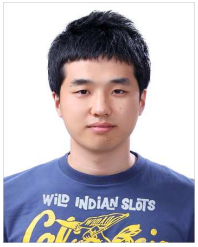
[7] D. d. O. Eduardo Ogasawara and P. Valduriez, "An algebraic approach for data-centric scientific workflows," in *Proceedings of the 37th International Conference on Very Large Data Bases*. VLDB Endowment, September 2011, pp. 619–626.

[8] e. a. Yolanda Gil, "Examining the challenges of scientific workflows," *IEEE Computer*, vol. 40, no. 12, pp. 24–32, December 2007.

[9] C. A. Ellis, "Information control nets: A mathematical model of information flow," in *Proceedings of Conference on Simulation, Modeling, and Measurement of Computer Systems*. ACM, 1979, pp. 225–240.

[10] K. P. Kim and C. A. Ellis, *Section II / Chapter VII. An ICN-based Workflow Model and Its Advances, Handbook of Research on BP Modeling*. IGI Global, ISR, 2009.

[11] V. Curcin and M. Ghanem, "Scientific workflow systems - can one size fit all?" in *Proceedings of the Cairo International Biomedical Engineering Conference*. IEEE, December 2008, pp. 1–9.



Hyun Ahn Hyun Ahn is a full-time Ph.D. student of computer science department and a graduate member of the collaboration technology research laboratory at Kyonggi University, South Korea. He received B.S. and M.S. degrees in computer science from Kyonggi University in 2010 and 2012, respectively. His research interests include workflow systems, BPM, scientific workflow systems, workflow-supported social and affiliation networks discovery, analysis, and visualization.

Minjae Park Minjae Park is a senior member of research staff at the solution R&D research center of BISTel, Inc., South Korea. He received B.S., M.S., and Ph.D. degrees in computer science from Kyonggi University in 2004, 2006, and 2009, respectively. His research interests include groupware, workflow systems, BPM, CSCW, collaboration theory, process warehousing and mining, workflow-supported social networks discovery and analysis, and process-aware factory automation systems.

Kwanghoon Pio Kim Kwanghoon Pio Kim is a full professor of computer science department and the founder and supervisor of the collaboration technology research laboratory at Kyonggi University, South Korea. He received B.S. degree in computer science from Kyonggi University in 1984. And he received M.S. degree in computer science from Chungang University in 1986. He also received his M.S. and Ph.D. degree from the computer science department of University of Colorado at Boulder, in 1994 and 1998, respectively.