Numerical Simulation and Analysis of Commercial Print Production Systems

Jun Zeng, I-Jong Lin, Eric Hoarau and Gary Dispoto; Hewlett-Packard Laboratories; Palo Alto, CA

Abstract

Digital transformation of commercial print production brings new opportunities including exploitation of embedded sensing and computing power that enable real-time communication during various phases of the production chain, and dynamic reconfiguration of production flow. Simulation based modeling can help to exploit these opportunities at both strategic and operational level. In this paper we report our ongoing work on simulating a print production system. We draw close parallel between print production design and electronic design automation, and model print production system as a network of interconnected, distinct processes. We describe our simulation framework, preliminary results, and validation against queueing network theory.

Introduction

The print production process has not yet been treated as a standard manufacturing activity due to its own unique characteristics: from the product's perspective, it simultaneously demands both product diversity and mass production; from investment's perspective, it is both capital-intensive and labor-intensive. The highly variable and dynamic job mix with highly personalized customer requirements results in many different combinations of equipment, resources, print shop configurations and business philosophy[1]. Many print shops rely on the mental models of one or a few skilled masters to tackle the complexities of the print production systems. Such artisan (or craftsman) decision-making practice has been identified as one of the key reasons that the productivity growth of the print industry as whole is falling far behind to other manufacturing industries[2].

The emergence of digital print production provides both challenges and opportunities to print production industry. The production agility enabled by digital print makes possible the integration of pervasive sensing and monitoring, real-time computing power, embedded actuation and reconfigurable workflow architecture, thus the transformation of the print production into an autonomous system formed by networked computing and physical components (*aka.* cyber-physical systems[3]). Print-on-demand pushes the product diversity to a new extreme: thousands of books can be printed within an hour and every page can be different. The content heterogeneousity of the digital print can be analogous to that of the Internet. A novel print production protocol, mirroring the TCP/IP Internet protocol, has been proposed that will create an Internet-like, highly efficient and robust print infrastructure and move the "lights-out" print factory closer towards reality[4].

To attack the low productivity problem associated with current print industry and to best uncover the full potential of the digital print promise, we propose a holistic, model-based approach that analyzes the design and management of print production as an integrated system, accounting for the performance, efficiency, stability, and sustainability as organic system attributes. This paper summarizes our ongoing effort towards this research direction. We map the analysis of print production system as an electronic design problem. We adapt an open-source electronic design automation (EDA) toolkit as our modeling platform. We model the print production as a heterogeneous, concurrent, integrated system.

In next section we describe the modeling problem and our approach. Following that we demonstrate system simulation results. In this section we also address validation issues. We conclude this paper by enumerating future works.

Problem Statement and Our Approach

The print production process [1] starts with jobs (customers' requests for prints) submitted through the store front. Once accepted, a unique *job ticket* is assigned. The submitted electronic files are examined for the correctness (*preflight*), edited for color quality and accuracy, imposed, and RIP-ed (*raster imaging process*). Printer-ready electronic files are sent to the printers to produce physical copies. In the case of book printing by web press, the printed web is first slit and cut into printed pages, which are then folded, collated, and bound into book blocks. The book blocks are joined by book covers, and then dust jackets. The finished books are sorted, labeled, and shipped, to conclude the purchase order.

Jobs arrive at a nondeterministic pace, in particular, when involving customers in proofing cycles. Other uncertainties associated with the jobs include the size, type (consequently workflow and equipment involved), and urgency. Variation associated with resources includes equipment capability, performance, stability, and costs of capital depreciation, material, energy, labor, etc. The variability in run-time policy includes job prioritization, timing, form and route associated with job release, possible monitoring and feedback, and scrap rate and quality assurance. The job profile is the input for system analysis and design. The design and analysis space is composed of the resource composition and the run-time policy engineering.

Here we would like to draw a close parallel between the analysis of print production and the electronic chip design (EDA) problems, in particular, that of labs-on-a-chip (LoC). LoC is a new system-on-chip architecture with primary focus on high-throughput, massively-parallel life science applications. A large number of independent assay operations can be carried out concurrently within this fingernail-sized chip; each involves a diverse set of sample operations including sample prep, assay, and detections. Similar to print production, a general-purpose LoC also anticipates content diversity of assay requests. The LoC synthesis techniques aim to provide solutions for scheduling assay operations, binding assay operations to available resources and

temporal intervals accounting for resource sharing and resource constraints, and module placement and route (layout)[5]. In print production systems, we face the same classes of design issues, about which, EDA provides extensive tools and methodologies in both formal methods and heuristics accumulated through 30 years of aggressive research.

We chose an open-source EDA toolkit, Ptolemy[6], as our modeling framework for print production systems. Ptolemy is a Java-based, actor-oriented modeling framework for concurrent, real-time, embedded systems. Ptolemy implements a set of well-defined models of computation (*e.g.*, continuous time, discrete event, finite state machine) that govern the communication among components. It provides a hierarchical component assembly design environment that enables the use of heterogeneous mixtures of models of computation (*e.g.*, hybrid and mixed-signal models). The print production system and control involves compute, logical and physical components, that, Ptolemy can blend together deploying different models of computation within the same simulation infrastructure.

We envision that the advanced state of our simulation infrastructure will integrate both the system components themselves and models of other components that are otherwise too difficult or too expensive to be included. Integrating the real components into simulation enables direct testing of the decision parameters. This approach benefits us the agility and cost-effectiveness that simulations bring and the higher level of fidelity that experiments provide. This simulation paradigm incorporates both hardware-in-the-loop and software-in-the-loop, enabling virtual-physical co-optimization. Ptolemy's strength in embedded systems design is an excellent fit. As a preliminary test, we have successfully constructed a Ptolemy model that integrates real-time raster imaging process as a component executed across intranet.

Simulation provides performance evaluation of a system design specified by a given set of design parameters. It is effective in discriminating among different design alternatives. However, simulation alone does not explicitly generate design parameters. For this, we develop an additional layer of models, referred to as the synthesis layer or generative models in Fig. 1. The generative models produce a set of design parameters according to the design objectives formalized by the objective function and constraints. Simulations are deployed to map the design parameters and the objective functions and to provide design verification.

EDA provides a theoretical foundation, suite of classical techniques, and well-established design tools that can be used for print production design. However, the unique characteristics of print production process require these EDA techniques be tailored for print production applications. Our research is aimed to provide bridge between EDA methodology and print production application. Fig. 1 illustrates this print production modeling environment that we aim to develop.

Research on productivity improvement is gaining momentum in print industry in recent years. Most notably is the LDP solution by Xerox[7]. LDP is Xerox's simulation-based service solution that aims to enhance print shop productivity. Similar to LDP, our work will also build upon simulations. However, we approach print production as content-driven cyber-physical systems; we anticipate our innovations on pervasive sensing and computing-based knowledge discovery will provide new means for monitoring and control thus larger design space and flexibility.

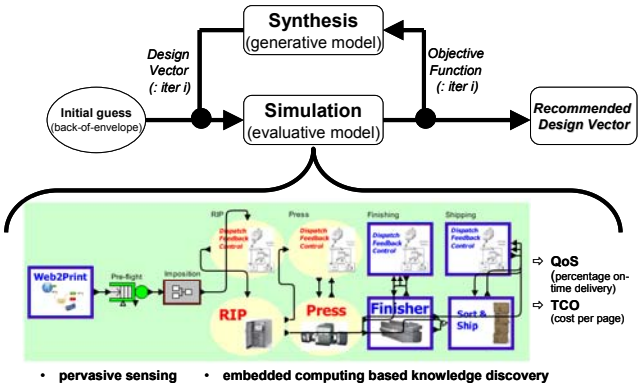


Figure 1. *Sketch illustrates our modeling infrastructure for print production system. It is composed of both simulation and synthesis layers. The simulation integrates both devices models (components with frames) and devices themselves (represented by frameless components).*

System Simulation

Fig. 2 illustrates the implementation of the simulation model of print production system using Ptolemy. The web-to-print *actor* (component class used in simulation) generates jobs used in simulation. Multiple event generators are used to trigger job generation (simulated as Poisson processes) concurrently, simulating the presence of multiple storefronts. A job is a freely-extensible record containing the job arrival time, due, shipping address, array of book records, etc.. Each book record corresponds to a unique book, containing number of pages, book sizes, paper types, and number of copies ordered. The web-to-print component uses random number generators of prescribed distributions to simulate the inherently stochastic nature of job description. Fig. 3a shows the statistical makeup of the jobs used in simulation. This record form of job description allows programming flexibility and straightforward translation into JDF. The use of job in simulation resembles the *job ticket* in print shops: the job travels through all

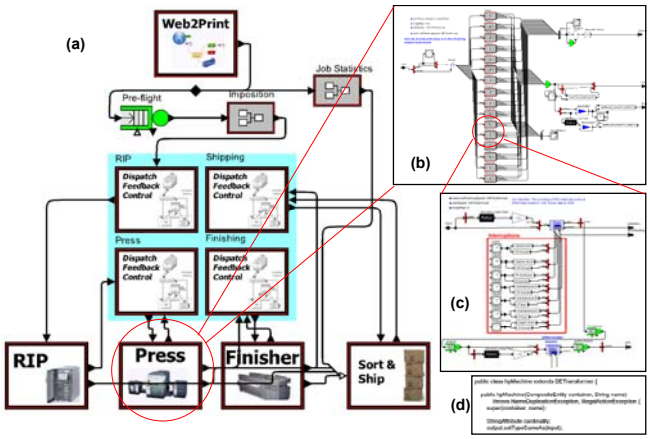


Figure 2. *An explosive view of the simulation model hierarchy of print production system. (a) is a screenshot of the top level of the simulation model illustrating the end-to-end print production process. (b) zooms in to show the implementation of the press zone where an array of presses generate printed pages. (c) further zooms in to show the implementation of a single press. (d) zooms in even further to show a custom implementation of an actor class created within the Ptolemy framework.*

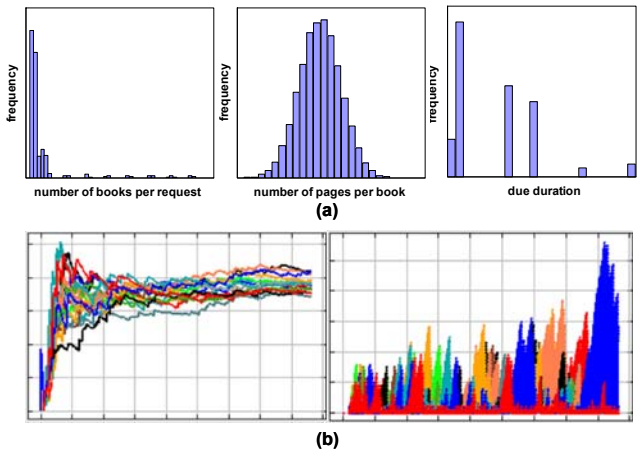


Figure 3. *Example simulation inputs and results. (a) shows the statistical makeup of the jobs used as the simulation input, extracted from 5,000 jobs (totally 150,000 books), from left to right, the distribution of number of books per job, number of pages per book, and the due duration of requests. (b) shows transient performance of 16 machines within the same process stage, from let to right, the utilization and number of jobs in waiting, the horizontal axis is time.*

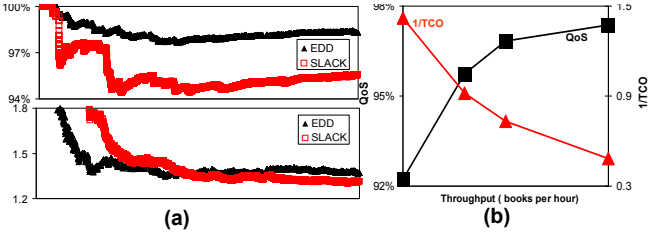
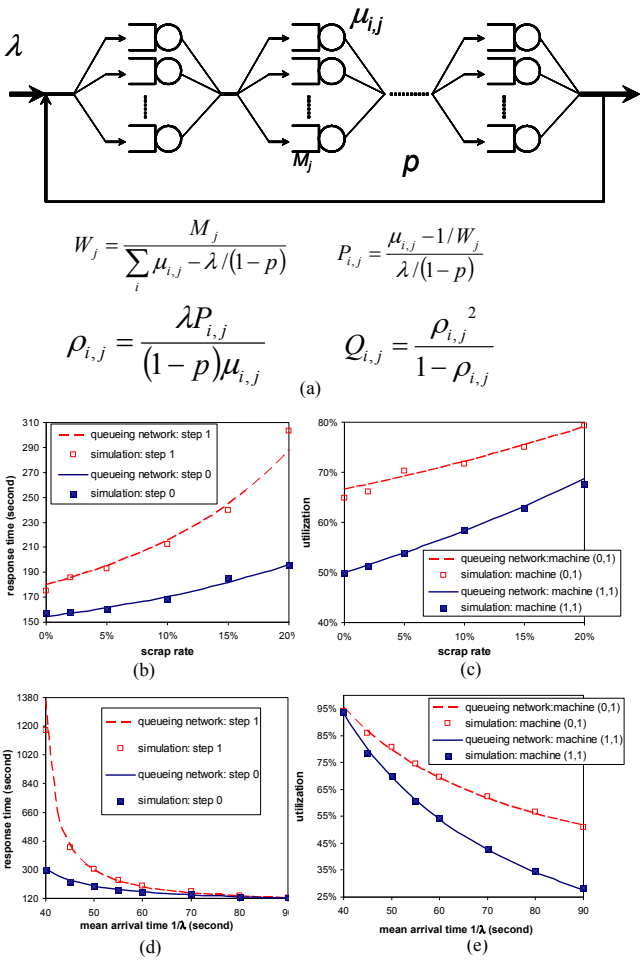


Figure 4. *Example system simulation results. (a) shows the impact of different job prioritization algorithms. Top shows QoS as percentage of on-time delivery. Bottom shows TCO as cent per page. The horizontal axis is time. (b) illustrates the relationship between the objective of maximizing QoS and the objective of minimizing TCO. The horizontal axis is throughput. The increase of throughput reduces TCO but may impact on-time delivery negatively as it may increase the response time.*

stages of print production process; each stage can read and/or write to it.

Jobs can be administered as first-come-first-serve basis. To ensure the on-time delivery and maximize the productivity, other heuristic *job priority policies* have been exploited, for instance, earliest-due-date, shortest-processing-time, minimum-slack-time (slack time is defined as the difference between the due date and the sum of the remaining work and the current time). As illustrated in Fig. 2b, usually there are array of machines that can perform the same task with various capabilities (*e.g.*, service time and capacity). A prioritized job needs to be assigned to a particular machine out of the machine array. Round-Robin with fixed-size quantum (*e.g.*, total page count) can be used. Alternatively, *machine priority policies* can be implemented, for instance, shortest-queue, shortest-response-time. In addition, to minimize the machine setup time batching can be used. The batch size needs to be engineered to minimize the negative impact on on-time delivery since urgent jobs may be delayed. Problems of this kind, constrained by the uncertainty and heterogeneority of the content and the availability of resources, have been proven NP-complete.



$$W_j = \frac{M_j}{\sum_i \mu_{i,j} - \lambda/(1-p)} \qquad P_{i,j} = \frac{\mu_{i,j} - 1/W_j}{\lambda/(1-p)}$$

$$\rho_{i,j} = \frac{\lambda P_{i,j}}{(1-p)\mu_{i,j}} \qquad Q_{i,j} = \frac{\rho_{i,j}^{\,2}}{1-\rho_{i,j}}$$

(a)

Figure 5. *Simulation validation by queueing network theory. (a) illustrates the validation problem. (b)-(e) shows the comparison between the simulation results and that of queueing network theory. In this particular simulation, N=2, M_0=3, and M_1=2. $1/\mu_{i,j}$ ={1, 1.5, 2; 1, 1.5} minutes. In (b) and (c), the mean inter-arrival time $1/\lambda$ is 1 minute and scrap rate varies from 0% to 20%. In (d) and (e), the scrap rate is 5% and $1/\lambda$ varies from 40 seconds to 90 seconds. Good agreement between simulation and queueing network theory is observed.*

The optimal heuristic practices are often print shop specific, depending on the characteristics of the content and resources.

The implementations of the machine *actors* account for the interruptions including regular maintenance and unexpected machine failure, as illustrated in Fig. 2c. Fig. 3b shows example simulation results of machine utilizations and queue lengths. They can be used to interrogate the effectiveness of the runtime policies.

At the sort-and-ship stage, each book is sorted according to its shipping address. Once all the books for particular print request are collected, we calculate quality-of-service (QoS), defined as percentage of on-time delivery of completed jobs, and total-cost-of-ownership (TCO), defined as cost per page. The primary reference we use on cost is [8]. Fig. 4a shows example results of QoS and TCO. Fig. 4b illustrates the two design objectives, minimizing TCO and maximizing QoS, may not be aligned and even competing against each other indicating design opportunity.

The fidelity of the simulations is measured by the closeness it resembles the reality. Validation (goodness of model assumptions) and verification (correctness of model implementations) are significant and integral components of model development effort. Depending on problems of study and the accessibility of the experimental data, we infer simulation fidelity from analytical results, experimental measurements, and/or intuitions of domain experts. Analytical results provide not only means for validation but also verification, however only a much simplified model of real-life problems can be solved analytically. Agreements to measurements of a real-life system provide highest confidence, however the accessibility and interpretation of experimental data is a challenge. We plan to present our simulation results comparing to print shop measurements in a near future. Below we discuss a validation study employing the queueing network theory.

Consider the fulfillment of a job includes N sequential steps (Fig. 5a). Jobs arrive with a mean arrival rate of λ. At each step j, multiple machines (M_j) can perform the same function with various capabilities (of mean service rate $\mu_{i,j}$). The probability that a job is assigned to one of the M_j machines is determined such that statistically jobs receive equal response time. Upon completion of the last step, defect jobs (of scrap rate p) are routed to the initial step to repeat the fulfillment process. Both job arrival and machine service are assumed Poisson processes.

Even though the process described above is of great simplification of the real print production process, it resembles its essence and the simulation involves several key *actors*. This simplified process can be solved analytically using queueing network theory[9]. It is chosen as a validation case. Fig. 5a shows the analytical solutions, including: W_j, response time mean; $P_{i,j}$, probability that an incoming job arriving at step j to be assigned to machine (i, j); $\rho_{i,j}$, utilization of machine (i, j); and $Q_{i,j}$, number of jobs accumulated at machine (i, j). Simulation results are plotted together with the analytical results shown in Fig. 5b-5e. The simulated mean values are extracted using sub-sampling method. Good agreements between the simulation results and that of the queueing network theory are observed.

Summary and Future Work

Simulation based modeling can help to optimize the print production system at both strategic and operational level. It can help to determine optimal workflow, predict production bottleneck, guide capacity planning, and even incubate completely new, disruptive print production paradigms without incurring material expenses. In this paper we report our ongoing work on simulating a print production system adopting EDA tools and methodology, and preliminary simulation results. After validating the simulations with print shop measurements, next step we plan to focus on developing the synthesis layer and its applications to the system design. The effectiveness of model-based analysis relies on winning the acceptance from the decision-makers. This paper summarizes our first step in demonstrating the fidelity and usefulness of simulation and modeling, and prompting the model-based design approach.

Acknowledgement

We thank prof. E. A. Lee (CHESS, UC Berkeley) and the Ptolemy team for the Ptolemy software and excellent support, in particular, C. Brooks, T. Feng and B. Rodiers. We thank M. Abergel (HP/Indigo), J. Lammens (HP/LFP), J. Rowson (HP/CTO), and R. Yancu (HP/Indigo) for valuable inputs. JZ thanks T. Cooney (HP/ITP) and L. Tully (HP/ITP) for the support of earlier exploration of this work.

References

[1] Kipphan, H. (Ed.), *Handbook of Print Media: Technologies and Production Methods*, Springer, 2001

[2] Uribe, J., *Print Productivity: A System Dynamics Approach*, M.S. thesis, Rochester Institute of Technology, 2007

[3] *National Science Foundation Cyber-Physical Systems Summit*, 2008.

[4] Lin, I-J., Zeng, J., Hoarau, E. and Dispoto, G., "Proposal for Next-Generation Commercial Print Infrastructure: Gutenberg-Landa TCP/IP", *NIP25*, 2009.

[5] Chakrabarty, K. and Su, F., *Digital Microfluidic Biochips: Synthesis, Testing, and Reconfiguration techniques*, CRC Press, 2006

[6] Brooks, C., Lee, E.A., Liu, X., Neuendorffer, S., Zhao, Y. and Zheng H., *Heterogeneous Concurrent Modeling and Design in Java*, University of California at Berkeley, 2008.

[7] Rai, S., Duke, C.B., Lowe, V., Quan-Trotter, C. and Scheermesser, T., "LDP Lean Document Production – OR-enhanced productivity improvements for the print industry", *Interfaces*, 39(1), 2009, 69-90.

[8] *Budgeted Hourly Cost Study: Digital Prepress Operations; Sheetfed Press Operations; Web Press Operations; Bindery, Finishing & Mailing Operations*. NAPL, 2008.

[9] Wolff, R.W., *Stochastic Modeling and The Theory of Queues*, Prentice Hall, 1989.

Author Biography

Jun Zeng is a senior scientist with Hewlett-Packard Laboratories. PhD (mechanical engineering) and MS (computer science) from Johns Hopkins University, and BS (modern mechanics) from USTC, China. Jun's publication includes 30 peer-reviewed papers and a co-edited book on CAD. He was a guest editor of IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems. Jun also serves as termed faculty and member of PhD dissertation committee with Duke University's Electrical and Computer Engineering Department.

I-Jong Lin is the research manager for GPU-RIP Print Services group with Hewlett-Packard Laboratories. He received PhD and MS in Electrical Engineering from Princeton University and Stanford University, respectively, and BS in Computer Systems Engineering from Stanford University. His research experiences span from VLSI-CAD, digital design, neural networks, video object extraction, to digital print. He is the author of "Video Object Extraction and Representation: Theory and Applications" by Kluwer Academic Publishers. I-Jong holds 13 patents.

Eric Hoarau is currently a senior researcher at Hewlett-Packard Laboratories in Palo Alto, CA. He received his B.S. from the University of California at Berkeley and his M.S. in Mechanical Engineering from the Massachusetts Institute of Technology. His research interests span several fields: Mechatronic systems, imaging systems, color imaging algorithms, distributed computing and system dynamics.

Gary Dispoto is the director of the Print Production Automation Lab, which seeks to streamline the processes required to produce industrial and commercial digital print, enabling new types of printed products and broader access to commercial print production. He holds several U.S. patents related to color imaging. Gary received B.S. and M.S. degrees in electrical engineering from Stanford University and an MBA degree from the University of Santa Clara.