R. Polig, K. Atasu, C. Hagleitner – IBM Research Zurich
L. Chiticariu, F. Reiss, H. Zhu – IBM Research Almaden
P. Hofstee – IBM Research Austin

Hardware-accelerated Text Analytics

Outline

- **Introduction & background**
- SystemT text analytics software
- Hardware-accelerated SystemT
- Experiments & conclusions

Source: If applicable, describe source origin

Text analytics

For years, Microsoft Corporation CEO Bill Gates was against open source. But today he appears to have changed his mind. “We can be open source. We love the concept of shared source,” said Bill Veghte, a Microsoft VP. “That's a super-important shift for us in terms of code access.”

Richard Stallman, founder of the Free Software Foundation, countered saying...

Name	Title	Organization
Bill Gates	CEO	Microsoft
Bill Veghte	VP	Microsoft
Richard Stallman	Founder	Free Software F...

Source: Example by Cohen, 2003

Big (text) Data

Press releases

Over **1.15 billion** Facebook users

Over **3 million** LinkedIn Company Pages

News posts

Blog posts

On average, over **400 million** tweets being sent per day

Machine log data

Scientific publications

Internal company data

Over **500,000** patent applications per year

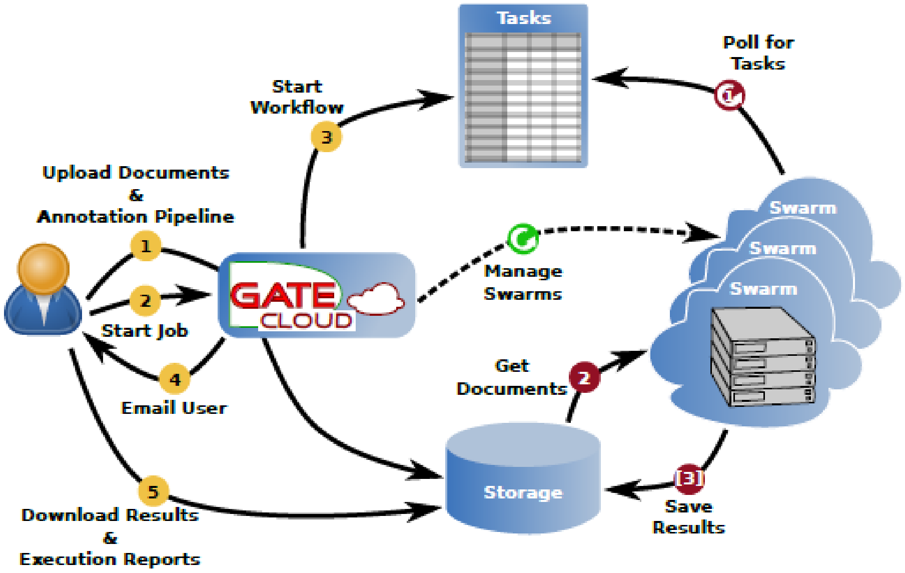
Source: www.socialmediatoday.com - statistics 2013
USPTO – Statistics

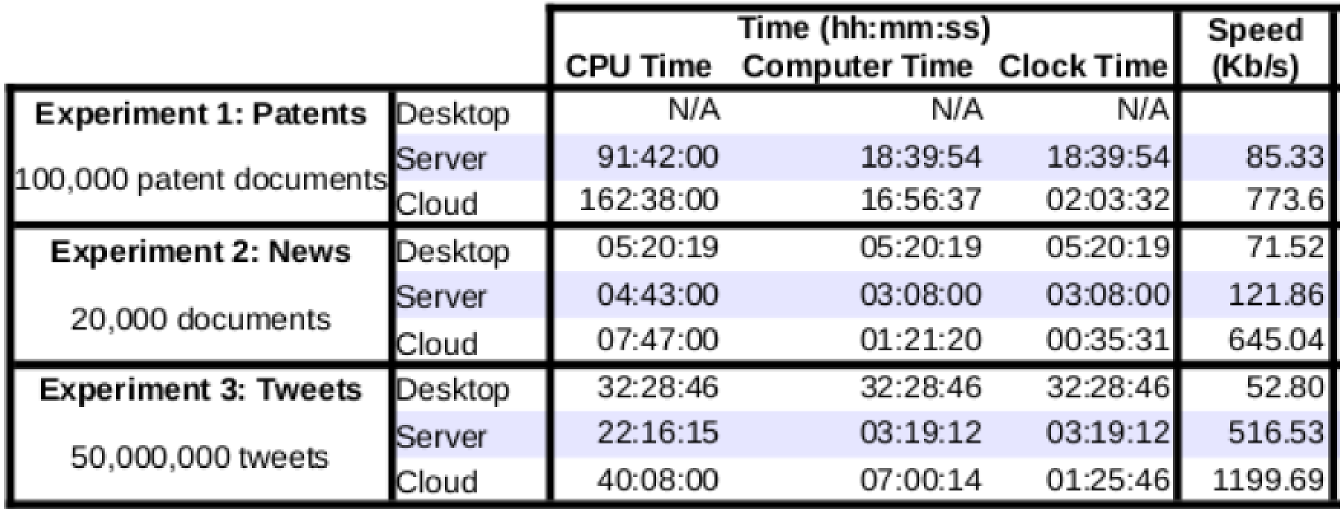
Performance limitations

- Throughput of information extraction systems is very limited
- Agatonovic et al. report ~200kB/s in 2008
- Analyzing 100GB of data required six days using an optimized IE system on 12 threads
  - USPTO DB is multiple TB
- CPUs are inefficient due to their limited parallelism for documents and operators

Source: M. Agatonovic, N. Aswani, K. Bontcheva, H. Cunningham, T. Heitz, Y. Li, I. Roberts, V. Tablan, Large-scale, parallel automatic patent annotation, in Proceedings of the 1st ACM workshop on Patent information retrieval, ACM, 2008, pp. 1.

State of the art – text analytics

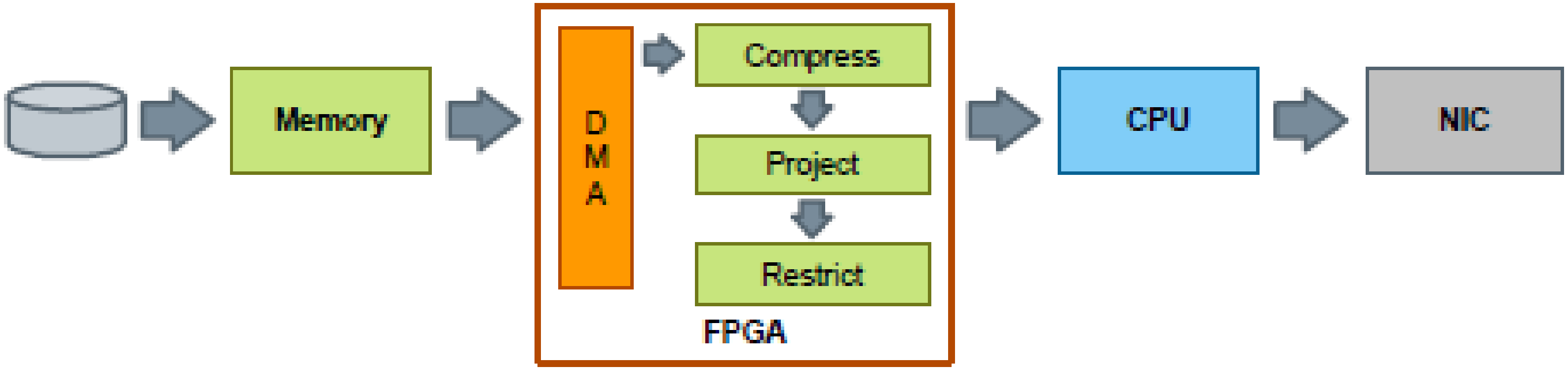




		Time (hh:mm:ss)			Speed (Kb/s)
		CPU Time	Computer Time	Clock Time	
Experiment 1: Patents 100,000 patent documents	Desktop	N/A	N/A	N/A	
	Server	91:42:00	18:39:54	18:39:54	85.33
	Cloud	162:38:00	16:56:37	02:03:32	773.6
Experiment 2: News 20,000 documents	Desktop	05:20:19	05:20:19	05:20:19	71.52
	Server	04:43:00	03:08:00	03:08:00	121.86
	Cloud	07:47:00	01:21:20	00:35:31	645.04
Experiment 3: Tweets 50,000,000 tweets	Desktop	32:28:46	32:28:46	32:28:46	52.80
	Server	22:16:15	03:19:12	03:19:12	516.53
	Cloud	40:08:00	07:00:14	01:25:46	1199.69

Source: V. Tablan, I. Roberts, H. Cunningham, K. Bontcheva, Gatecloud. net: a platform for large-scale, open-source text processing on the cloud, Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences 371, 2011

State of the art – Query compilation



NETEZZA – FAST Engine

- Query compilation for FPGAs has seen high interest in recent years
- Netezza applicances accelerate SQL queries by using the FPGA as a pre-filter when reading data from disk
- Quick query compilation and generation can be achieved by using dynamic partial reconfiguration
- All compiled designs miss complex operations such as regular expression matching and joins

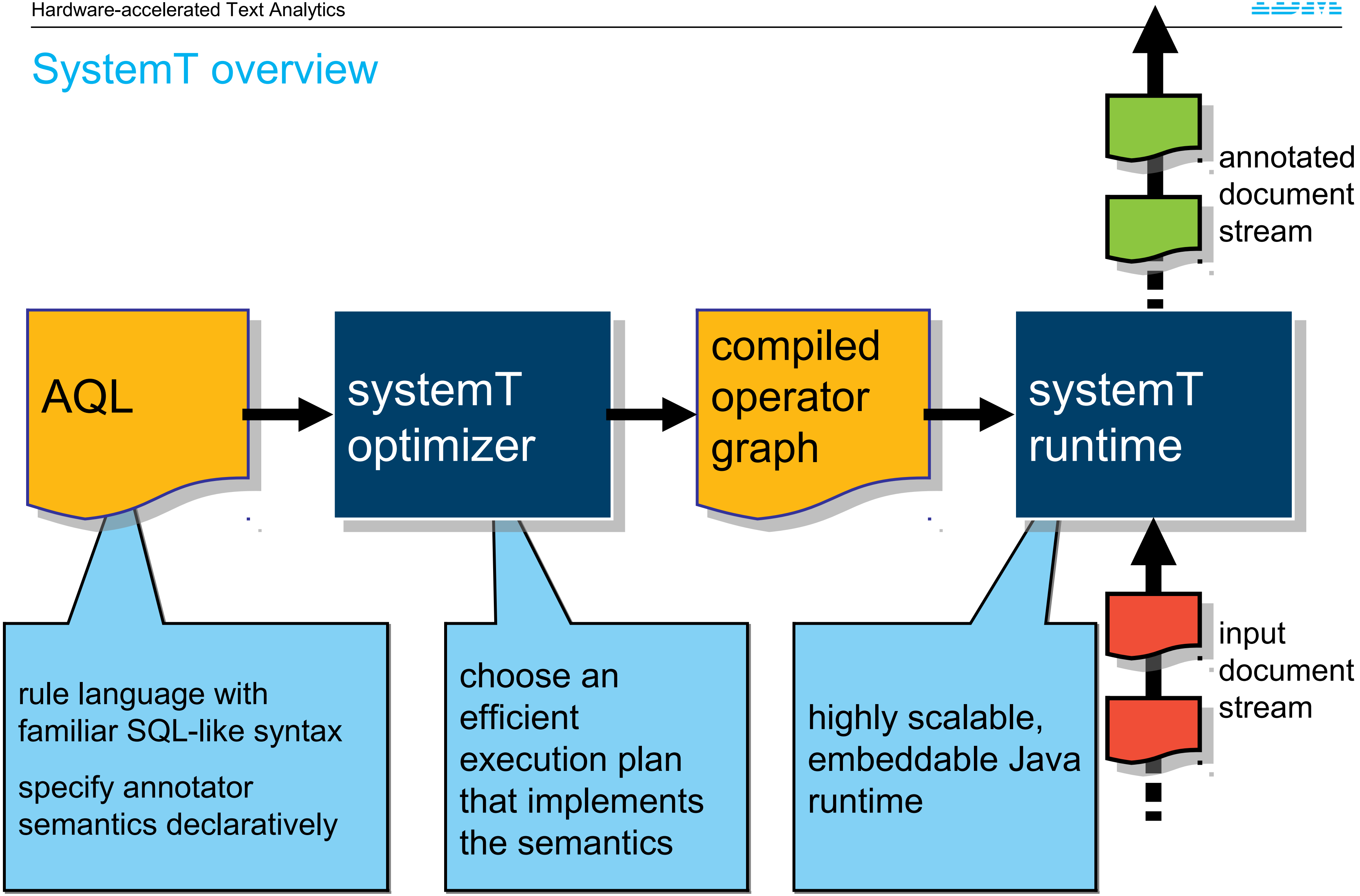
- C. Dennl, D. Ziener, J. Teich, On-the-fly composition of fpga-based sql query accelerators using a partially recongurable module library, in: Field-Programmable Custom Computing Machines (FCCM), 2012 IEEE 20th Annual International Symposium on, IEEE, 2012, pp. 4
- R. Mueller, J. Teubner, G. Alonso, Streams on wires: a query compiler for fpgas, Proceedings of the VLDB Endowment 2009

Outline

- Introduction & background
- **SystemT text analytics software**
- Hardware-accelerated SystemT
- Experiments & conclusions

Source: If applicable, describe source origin

SystemT overview



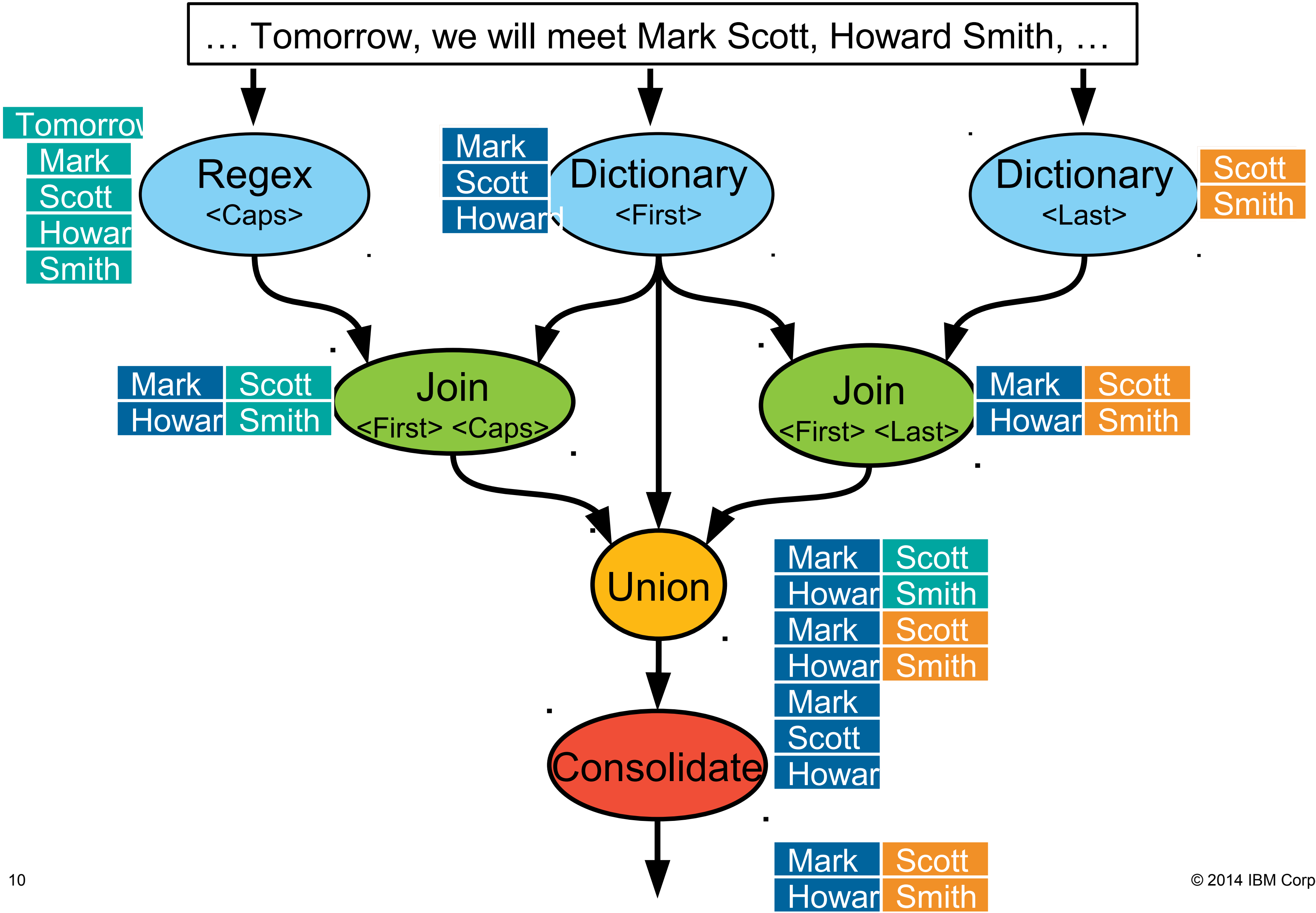
rule language with familiar SQL-like syntax

specify annotator semantics declaratively

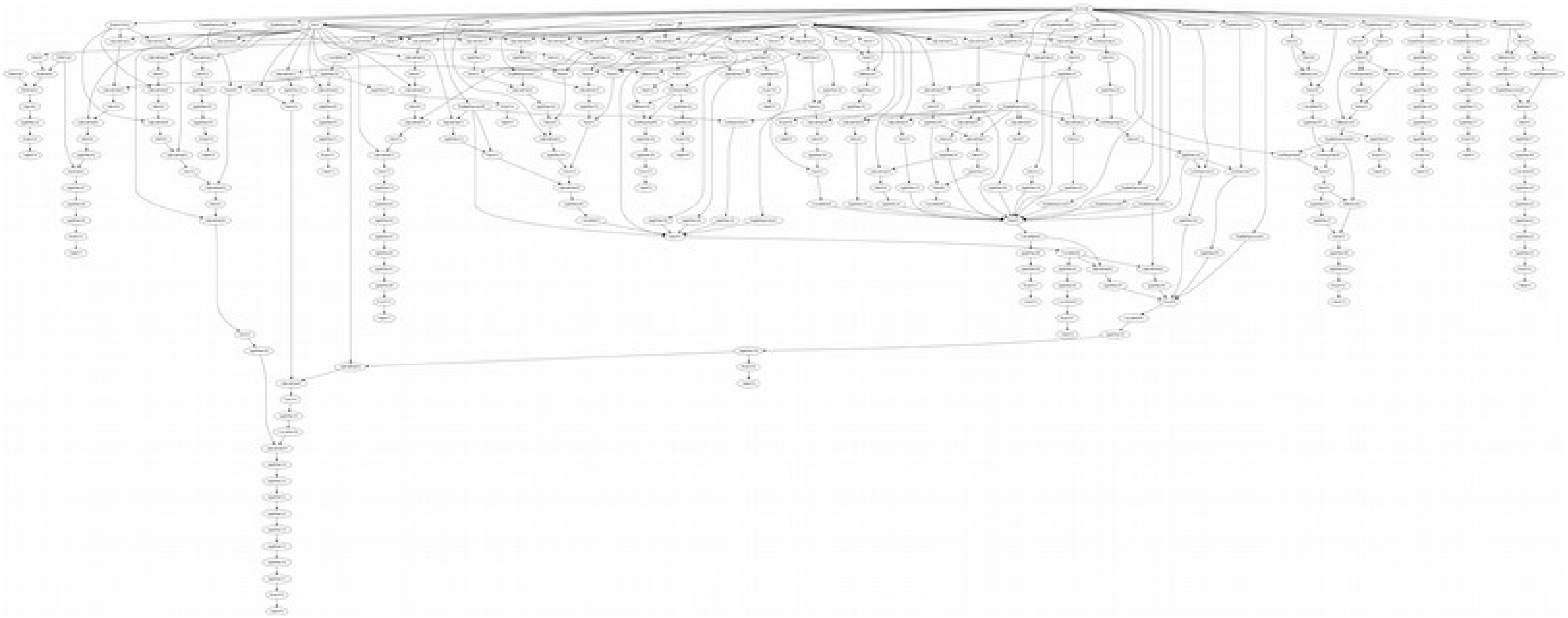
choose an efficient execution plan that implements the semantics

highly scalable, embeddable Java runtime

Annotation Operator Graph (AOG)



AOG of a real-life SystemT IE query

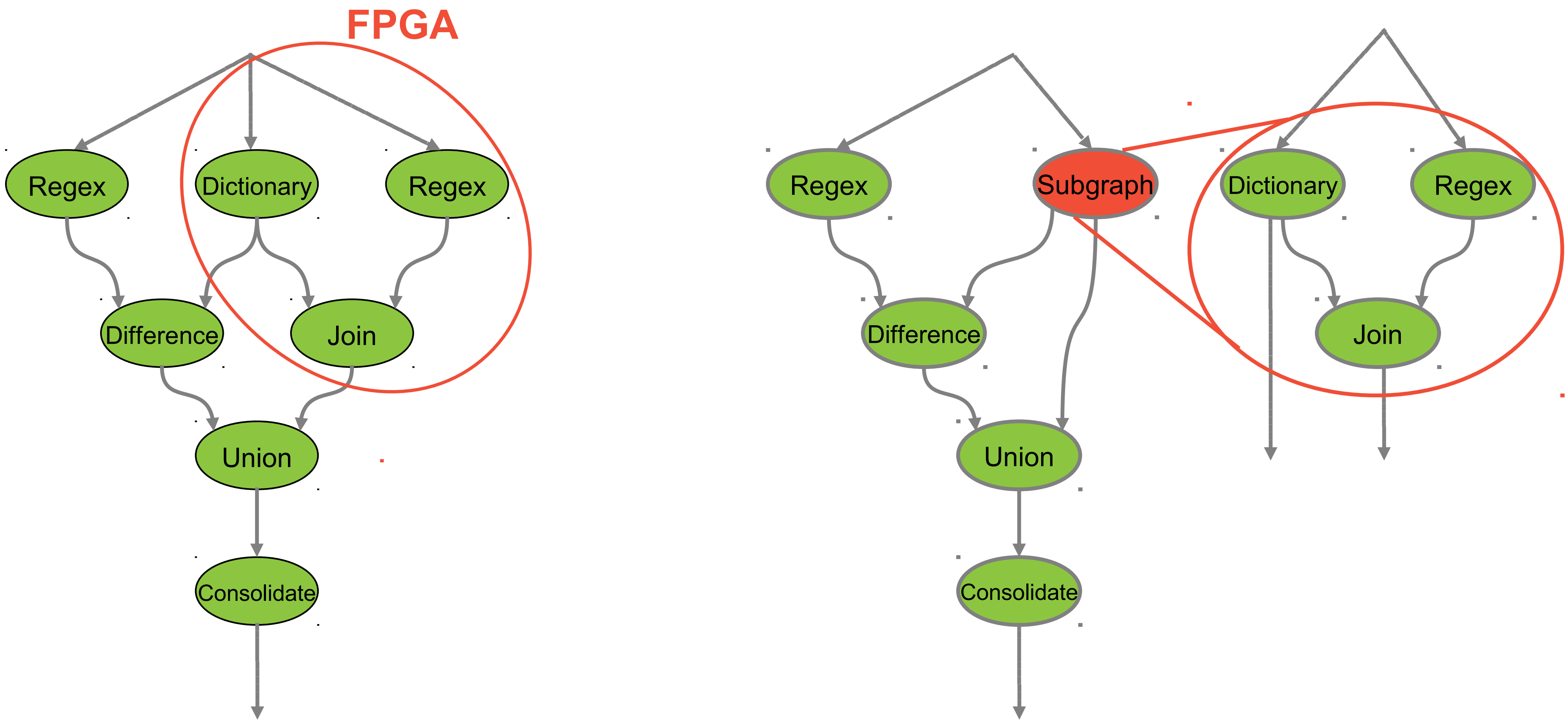


Outline

- Introduction & background
- SystemT text analytics software
- **Hardware-accelerated SystemT**
- Experiments & conclusions

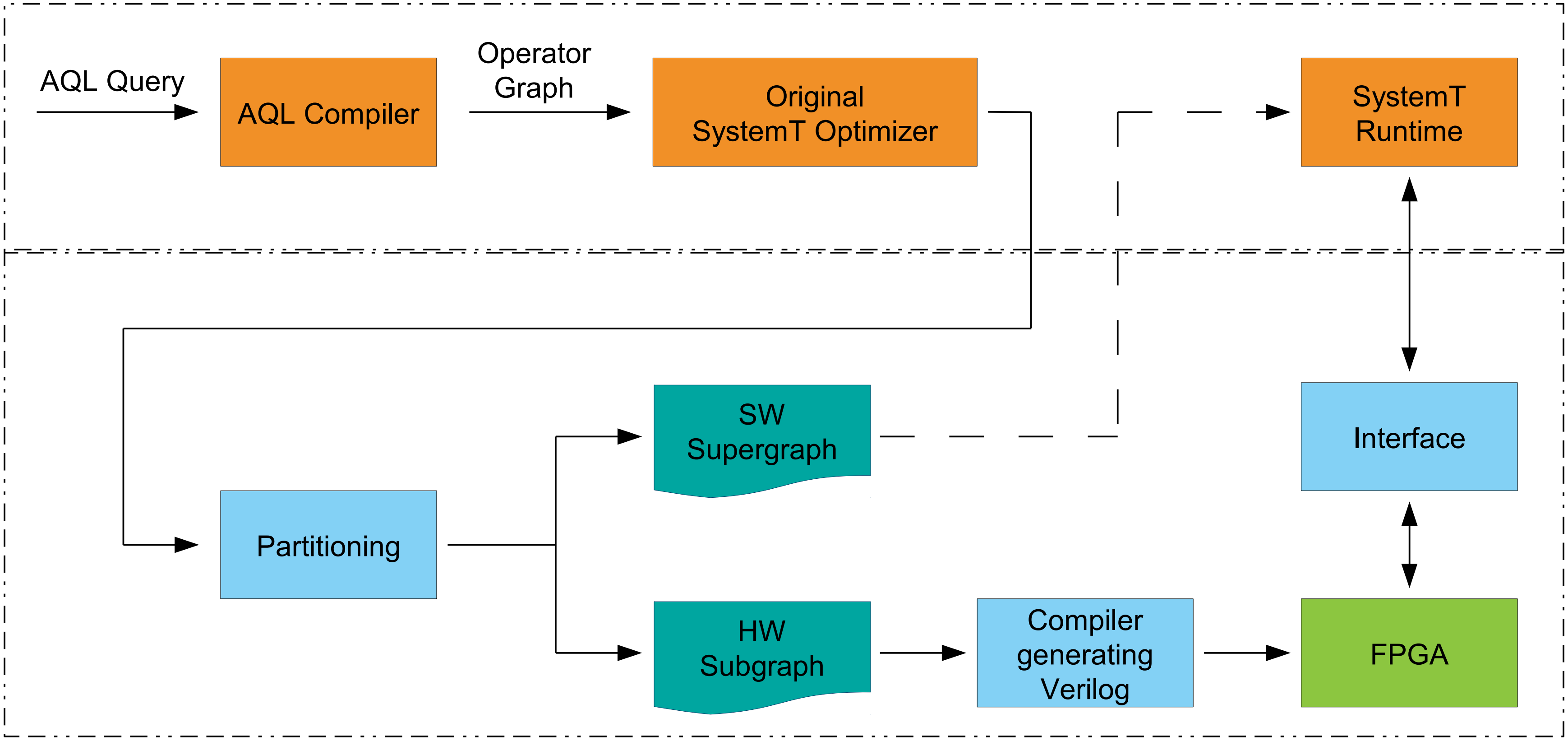
Source: If applicable, describe source origin

Acceleration concept



- Select subgraphs to run in HW
- Compile subgraphs for HW (FPGA)
- Generate interface for the custom HW
- Maximize throughput of the overall system

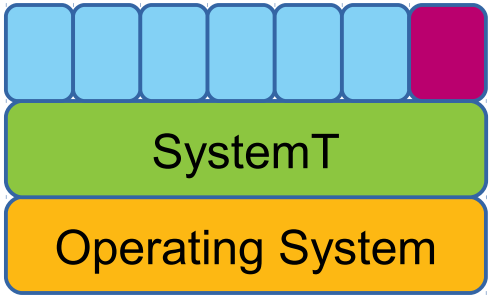
A hardware-accelerated text analytics system



Source: If applicable, describe source origin

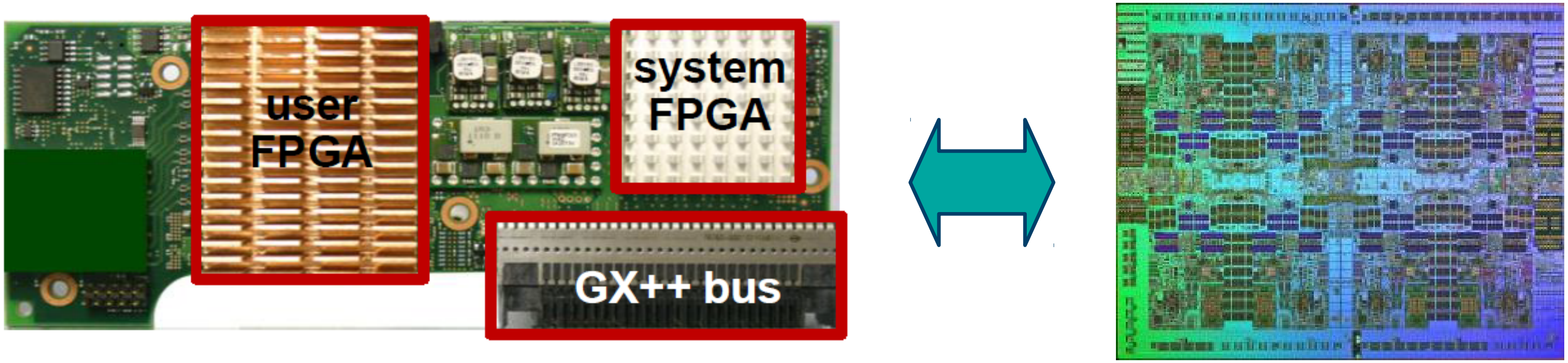
Deployment system

- CAPI predecessor system
- Software based MMU
- Virtual addressing from user FPGA



SystemT

Operating System

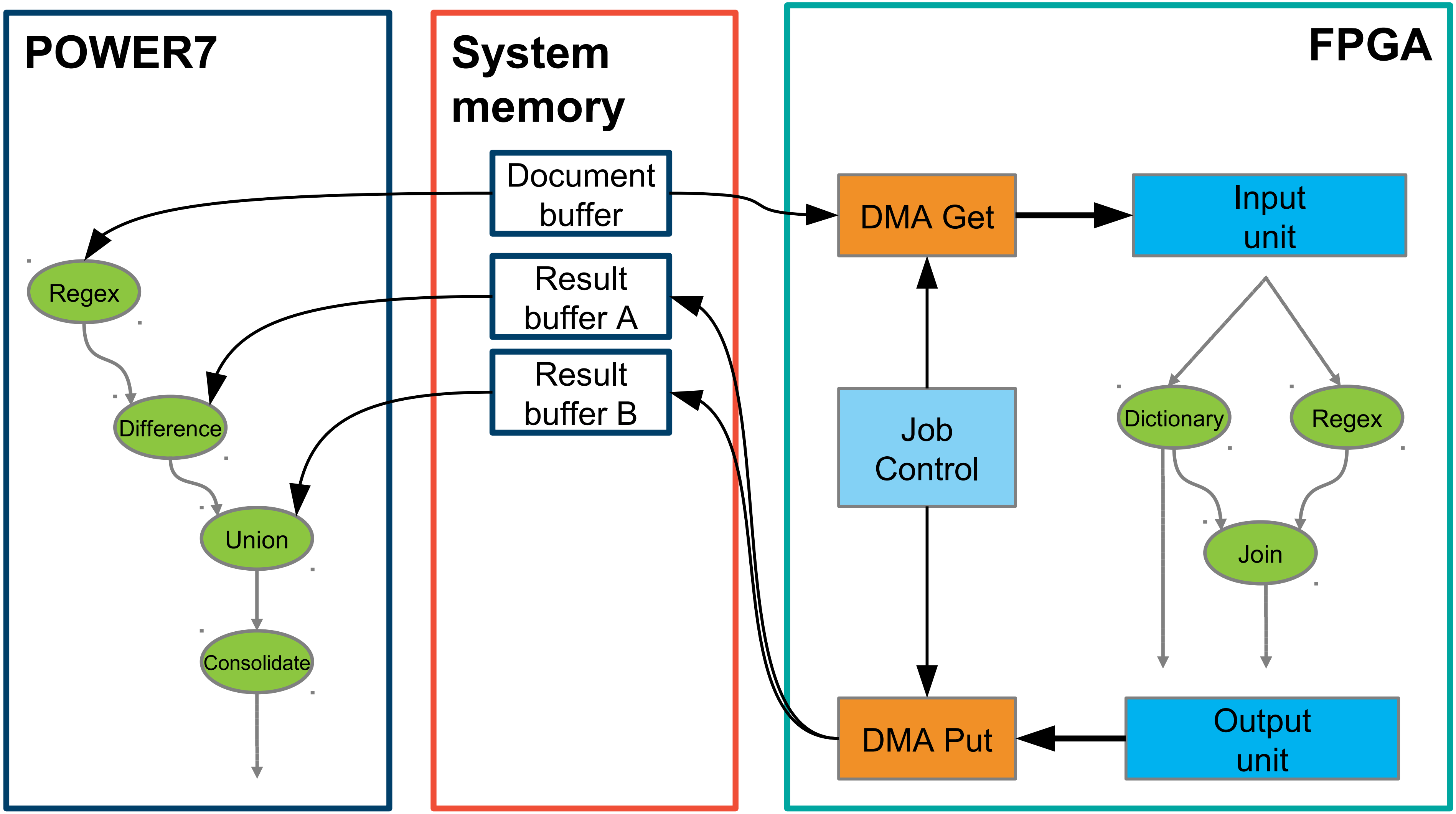


Stratix IV GX530

POWER 7

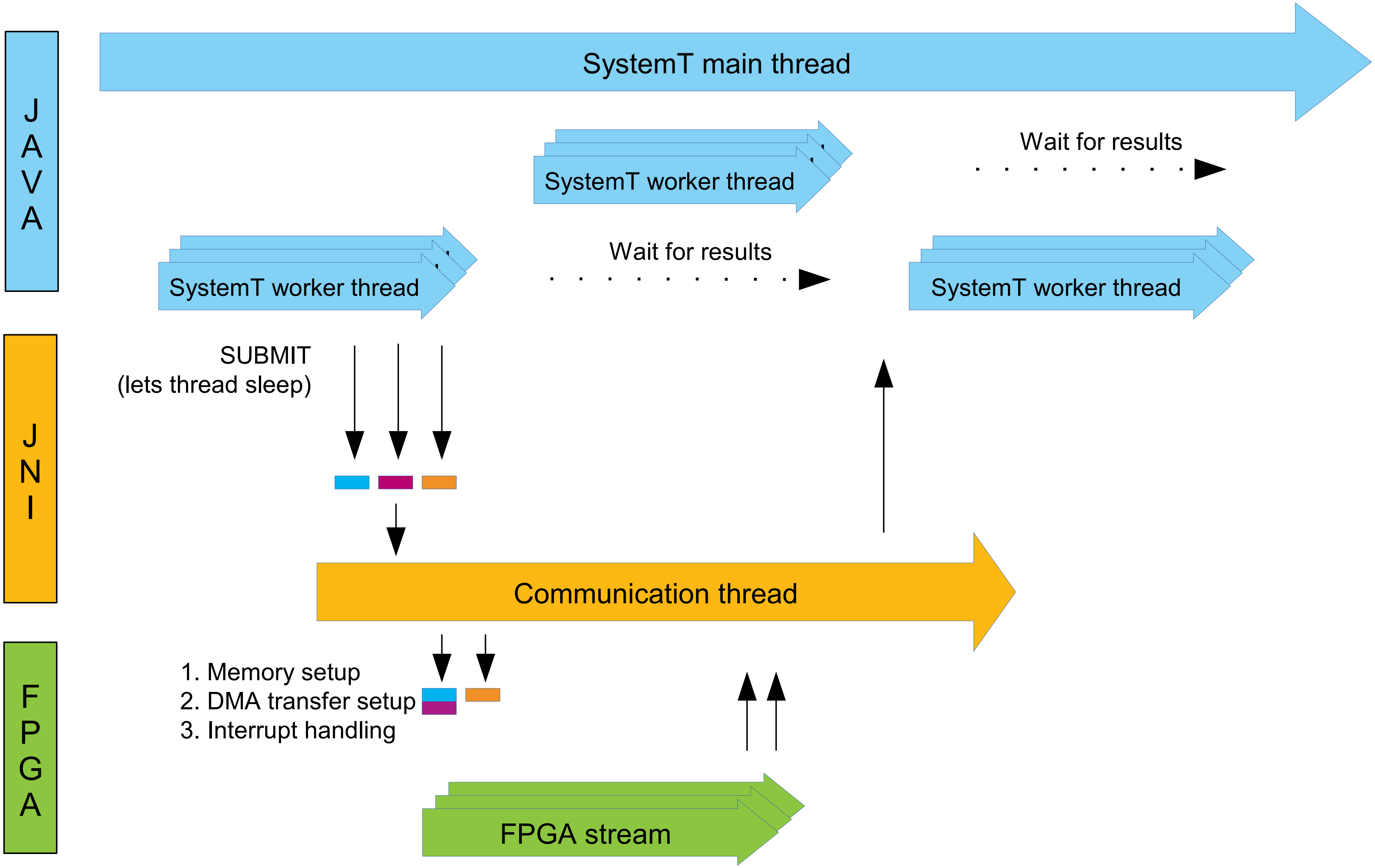
Source: If applicable, describe source origin

SystemT acceleration: Partitioned system



Source: If applicable, describe source origin

Communication scheme

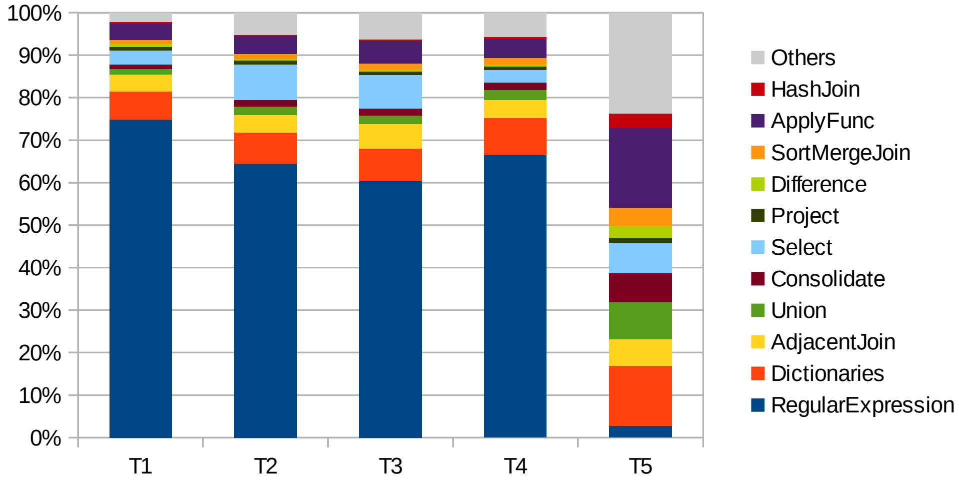


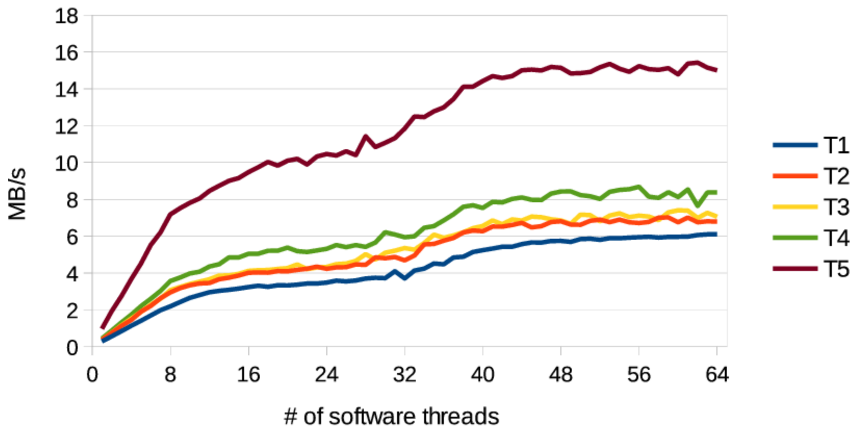
Outline

- Introduction & background
- SystemT text analytics software
- Hardware-accelerated SystemT
- **Experiments & conclusions**

Source: If applicable, describe source origin

Software results for five real-life information extraction queries

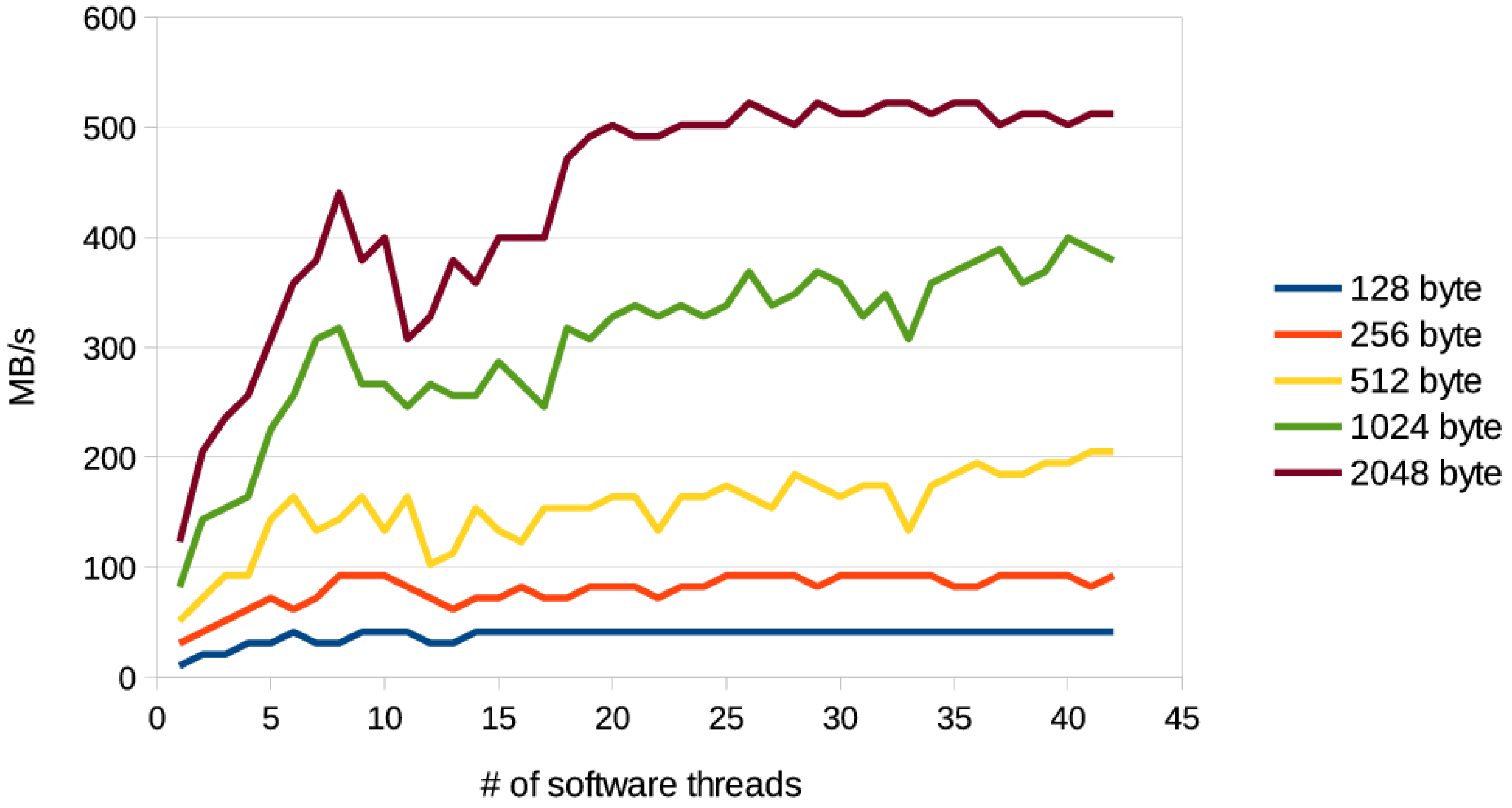




- Measurements on a two socket POWER7 server with 8 cores per CPU @3.55GHz

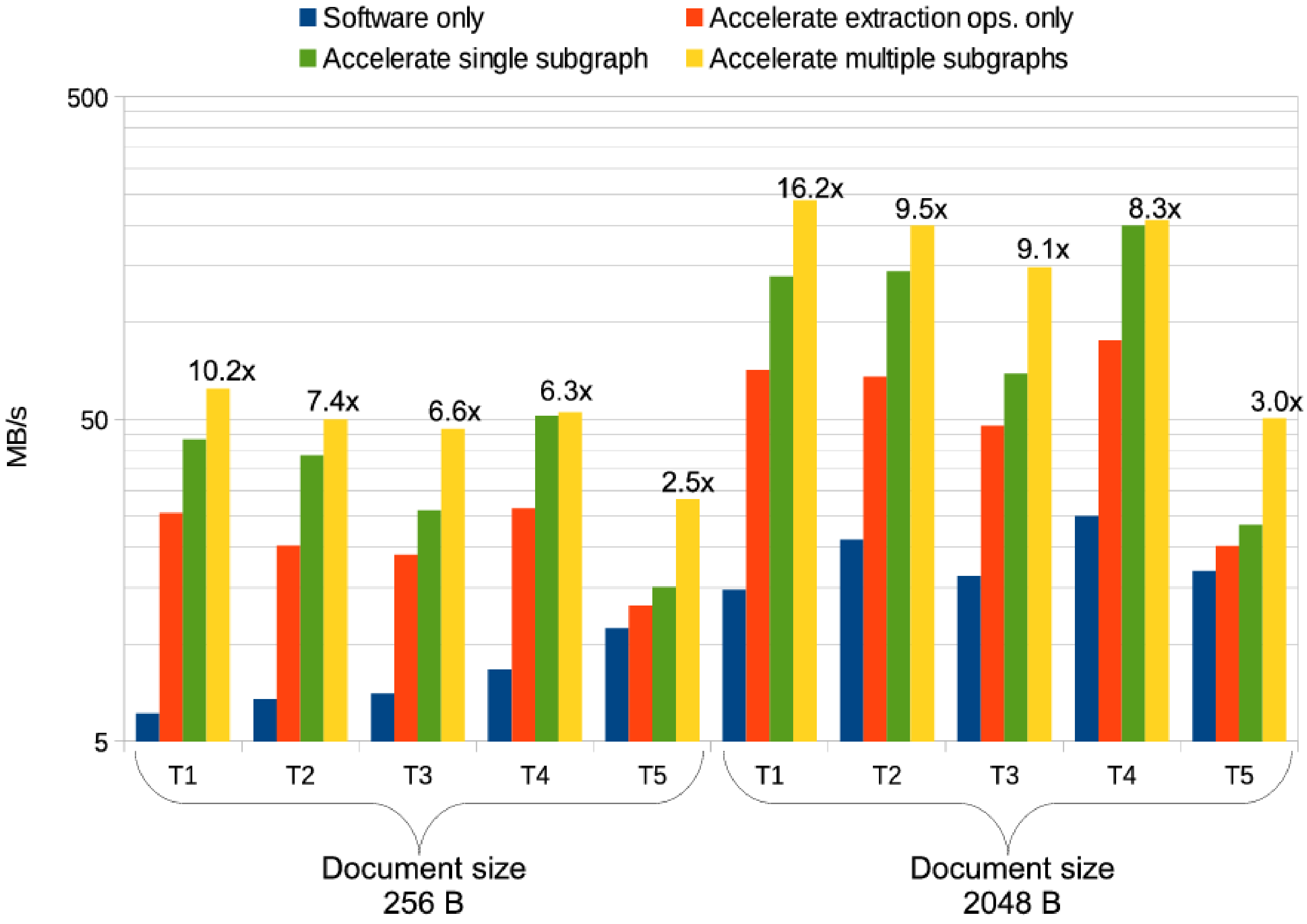
Measured system throughput rates

- @ 250 MHz, using 4 hardware threads, 125 MB/s per hardware thread → 500 MB/s
- Document sizes ranging from 128 bytes (twitter feeds) to 2048 bytes (news entries)



Speedup estimations

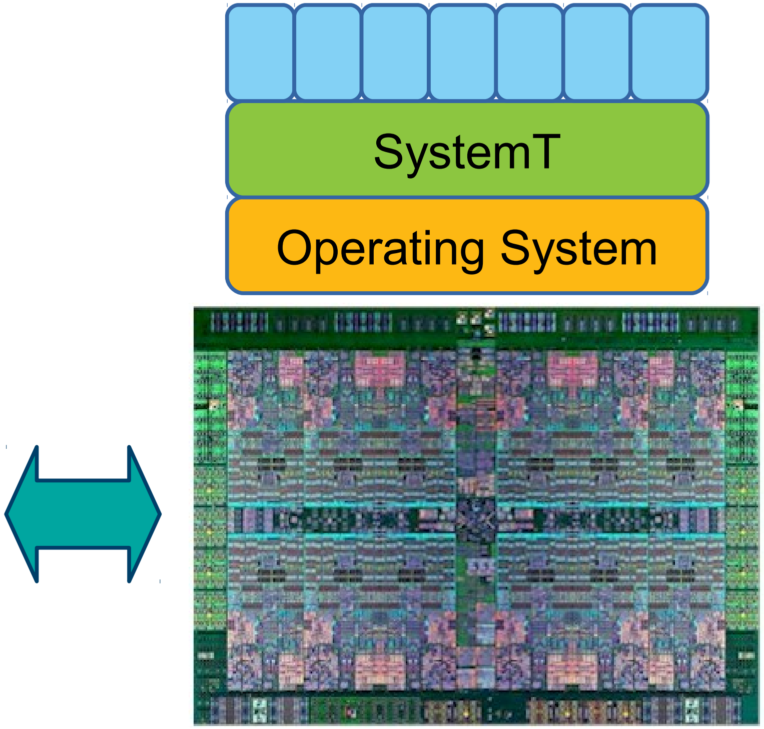
- Using 4 hardware threads → 500 MB/s
- Using 64 software threads on P7



Moving on to POWER8

- Higher SW performance
- CAPI system
- Virtual addressing from user FPGA
- Multiple cards per CPU

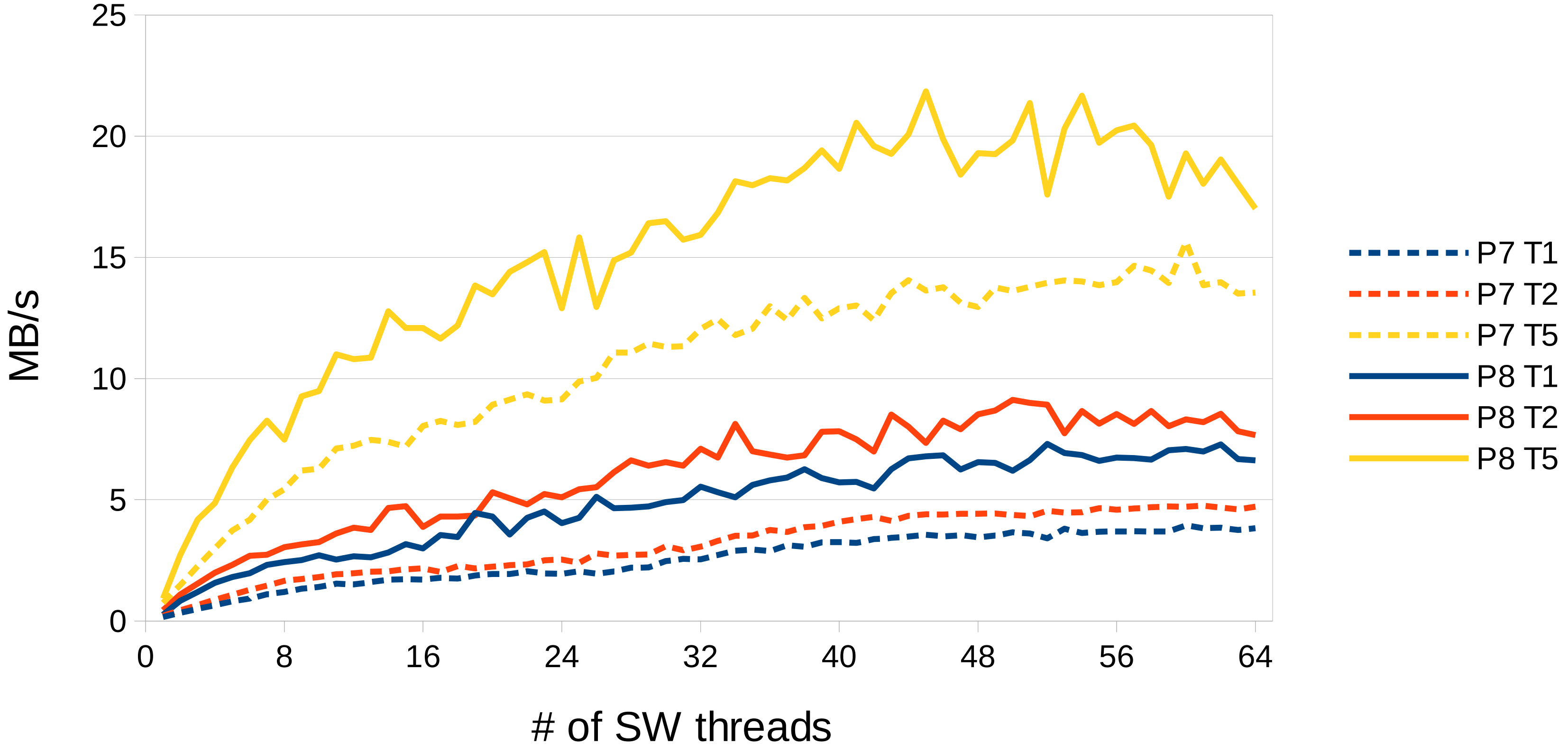
Stratix V GX A7



POWER 8

Source: If applicable, describe source origin

Software performance on POWER8



- POWER7 @ 3.55GHz
- POWER8 @ 3.8GHz

Source: If applicable, describe source origin

QUESTIONS?

Please feel free to contact us offline: {pol,kat,hle}@zurich.ibm.com

Acknowledgements

Andrew. K. Martin – IBM Research Austin
Kanak. B. Agarwal - IBM Research Austin
Eva Sitaridi – Columbia University