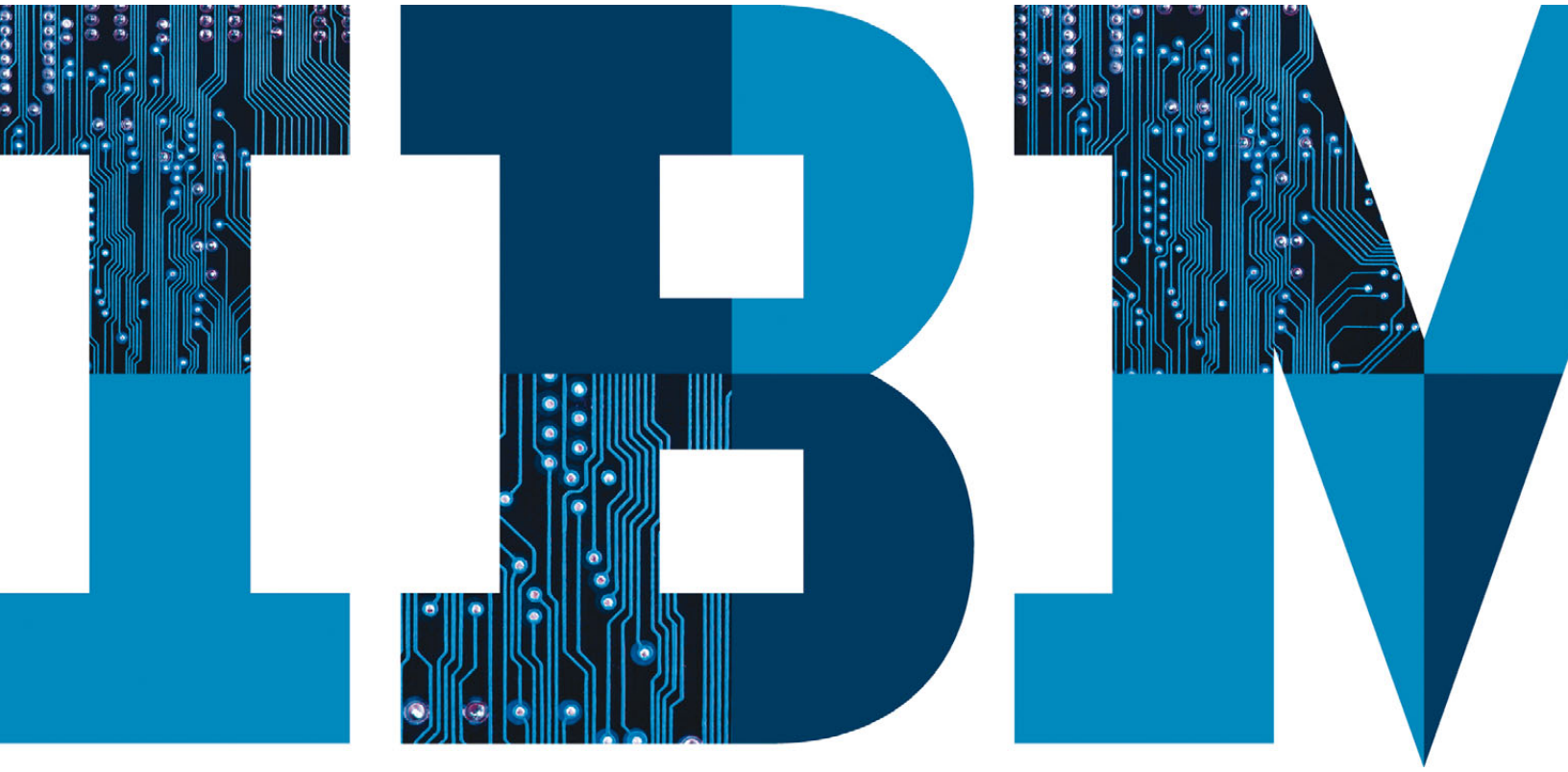
IBM Systems and Technology
An IBM White Paper

February 2011

Watson – A System Designed for Answers

The future of workload optimized systems design



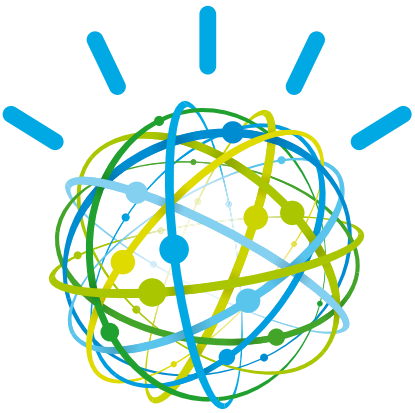
Executive summary

Over the last century, IBM has achieved numerous scientific breakthroughs through its commitment to research and its tradition of Grand Challenges. These Grand Challenges—such as Deep Blue®, which was designed to rival world chess champion Gary Kasparov—work to push science in ways that weren't thought possible before. Watson is the latest IBM Research Grand Challenge, designed to further the science of natural language processing through advances in question and answer technology.

Watson is a workload optimized system based on IBM DeepQA architecture running on a cluster of IBM® POWER7® processor-based servers. After four years of intense research and development by a team of IBM researchers, Watson competed on Jeopardy! in February 2011, performing at the level of human experts in terms of precision, confidence and speed against two of the best-known and most successful Jeopardy! Champions, Ken Jennings and Brad Rutter. This white paper explains Watson's workload optimized system design, how it's emblematic of the future of systems design, and why this represents a new computing paradigm.

Jeopardy! The IBM challenge

In 1997, Deep Blue, the computer chess-playing system developed by IBM Research, captured worldwide attention by competing successfully against world chess champion Gary Kasparov. It was the culmination of a grand challenge to advance the science of computing in a way that created great popular interest.



Today, with companies increasingly capturing critical business information in natural language documentation, there is growing interest in workload optimized systems that deeply analyze the content of natural language questions to answer those questions with precision. Advances in question answering (QA) technology will increasingly help support professionals in critical and timely decision making in areas such as health care, business intelligence, knowledge discovery, enterprise knowledge management, and customer support.

With QA in mind, IBM settled on a challenge to build a computer system called "Watson" (after Thomas J. Watson, the founder of IBM), which could compete at the human champion level in real time on the American TV quiz show *Jeopardy!* The program, which has been broadcast in the United States for more than 25 years, pits three human contestants against one another to answer rich natural language questions over a broad

range of topics, with penalties for wrong answers. In this three-person competition, confidence, precision and answering speed are of critical importance, as contestants usually come up with their answers in the few seconds it takes for the host to read a clue. To compete in this game at human-champion levels, a computer system would need to answer roughly 70 percent of the questions asked with greater than 80 percent precision in three seconds or less.

Watson represents an impressive leap forward in systems design and analytics. It runs IBM's DeepQA technology, a new kind of analytics capability that can perform thousands of simultaneous tasks in seconds to provide precise answers to questions. Powered by IBM POWER7 processor technology, Watson is an example of the complex analytics workloads that are becoming increasingly common and critical to business success and competitiveness in today's data-intensive environment.

Watson competed against two of the most well-known and successful *Jeopardy!* champions—Ken Jennings and Brad Rutter—in a two-match contest aired over three consecutive nights beginning on February 14, 2011.

IBM DeepQA

DeepQA is a massively parallel probabilistic evidence-based architecture. For the *Jeopardy!* Challenge, more than 100 different techniques are used to analyze natural language, identify sources, find and generate hypotheses, find and score evidence, and merge and rank hypotheses. Far more important than any particular technique is the way all these techniques are combined in DeepQA such that overlapping approaches can bring their strengths to bear and contribute to improvements in accuracy, confidence, or speed.

DeepQA is an architecture with an accompanying methodology, but it is not specific to the *Jeopardy!* Challenge. IBM has begun adapting it to different business applications and additional exploratory challenge problems including medicine, enterprise search and gaming.

The overarching principles in DeepQA are:

1. **Massive parallelism**: Exploit massive parallelism in the consideration of multiple interpretations and hypotheses.
2. **Many experts**: Facilitate the integration, application and contextual evaluation of a wide range of loosely coupled probabilistic question and content analytics.
3. **Pervasive confidence estimation**: No single component commits to an answer; all components produce features and associated confidences, scoring different question and content interpretations. An underlying confidence processing substrate learns how to stack and combine the scores.
4. **Integrate shallow and deep knowledge**: Balance the use of strict semantics and shallow semantics, leveraging many loosely formed ontologies.

Speed and scale-out

DeepQA is developed using Apache UIMA, a framework implementation of the Unstructured Information Management Architecture. UIMA was designed to support interoperability and scale-out of text and multimodal analysis applications. All of the components in DeepQA are implemented as *UIMA annotators*. These are components that analyze text and produce *annotations* or assertions about the text. Over time Watson has evolved so that the system now has hundred of components. UIMA facilitated rapid component integration, testing and evaluation.

Early implementations of Watson ran on a single processor, which required two hours to answer a single question. The DeepQA computation is embarrassing parallel, however, and so it can be divided into a number of independent parts, each of which can be executed by a separate processor. UIMA-AS, part of Apache UIMA, enables the scale-out of UIMA applications using asynchronous messaging. Watson uses UIMA-AS to scale out across 2,880 POWER7 cores in a cluster of 90 IBM Power® 750 servers. UIMA_AS manages all of the inter-process communication using the open JMS standard. The UIMA-AS deployment on POWER7 enabled Watson to deliver answers in one to six seconds.

Watson has roughly 200 million pages of natural language content (equivalent to reading 1 million books). Watson uses the Apache Hadoop framework to facilitate preprocessing the large volume of data in order to create in-memory datasets used at run-time. Watson's DeepQA UIMA annotators were deployed as mappers in the Hadoop map-reduce framework, which distributed them across processors in the cluster. Hadoop contributes to optimal CPU utilization and also provides convenient tools for deploying, managing, and monitoring the data analysis process.

Harnessing POWER7

Watson harnesses the massive parallel processing performance of its POWER7 processors to execute its thousands of DeepQA tasks simultaneously on individual processor cores. Each of Watson's 90 clustered IBM Power 750 servers features 32 POWER7 cores running at 3.55 GHz. Running the Linux® operating system, the servers are housed in 10 racks along with associated I/O nodes and communications hubs. The system has a combined total of 16 Terabytes of memory and can operate at over 80 Teraflops (trillions of operations per second).

With its innovative, eight-core processor design, POWER7 is ideally suited for massively parallel processing of Watson's analytics algorithms. POWER7 also features 500 gigabytes of on-chip communications bandwidth, contributing to exceptional efficiency of both memory and processor utilization. And since each server packs 32 high performance POWER7 cores with up to 512 GB of memory, the Power 750 makes an ideal platform for Watson's processor *and* memory-hungry Java processes.

Designing Watson on commercially available Power 750 servers was a deliberate choice to ensure more rapid adoption of optimized systems in industries such as healthcare and financial services. That goal was a fundamental difference between Watson and Deep Blue, which was a highly customized supercomputer. Deep Blue was based on an earlier generation of Power processor technology, featuring a 30 node RS/6000 SP system, with each node containing a single 120 MHz POWER2 processor. But in addition to the regular POWER2 processors, Deep Blue's performance was enhanced with 480 special purpose chess processor chips.

The same Power 750 server used by Watson is already deployed today by thousands of organizations in optimized systems that provide for both complex analytics and transaction processing. Rice University in Houston, Texas, for example, uses IBM Power 750 systems to accelerate the understanding of the molecular basis of cancer through the application of genome analysis technologies. POWER7 systems have given Rice more flexibility and efficiency, enabling them to pursue a broader range of research challenges on a single system than was possible before. And GHY International, a customs brokerage firm in Canada, migrated to a new Power 750 running Power AIX®, Power i and Power Linux to better support their clients' increased engagement in international trading. With PowerVM™ virtualization, GHY is now able to deploy new capabilities in as little as five minutes to support their clients' changing needs.

A system designed for answers

After four years of intense research and development by a team of IBM researchers, Watson has demonstrated its ability to compete on *Jeopardy!* against champion players, performing at human-expert levels in terms of precision, confidence and speed. The project has advanced the fields of unstructured data analytics, natural language processing, and the design of workload optimized systems. Beyond Jeopardy!, the technology behind Watson can be adapted to solve business and societal problems—for example, diagnosing disease, handling online technical support questions, and parsing vast tracts of legal documents—and to drive progress across industries.

Watson's ability to understand the meaning and context of human language, and rapidly process information to find precise answers to complex questions, holds enormous potential to transform how computers can help people accomplish tasks in business and their personal lives.

For more information

To learn more about Watson, POWER7 and workload optimized systems, please contact your IBM marketing representative or IBM Business Partner, or visit the following websites:

- **ibm.com**/systems/power/advantages/watson
- **ibm.com**/systems/power

IBM Systems and Technology Group
Route 100
Somers, NY 10589

Produced in the United States of America
February 2011

Please Recycle

POW03061-USEN-00