Introduction to Pattern Discovery through Data Qualification

Curt Harris, Managing Member
curt.harris@kennentech.com

Jack Ring, Member
jack.ring@kennentech.com

If you must discover whether any source data qualifies with respect to any part of a complex, combinatorial information topology then the General Purpose Set Theoretic Processor, GPSTP®, will make a breakthrough difference. Further, GPSTP® concepts may trigger new ideas for helping users navigate their ever-larger, future "meaning mazes."

The GPSTP® is a user-configurable nondeterministic finite state machine. Specifics of the architecture and method are described in US Patents #7,392,229B2, #7,487,131, #7,774,286 and #8,065,249.

This breakthrough capability opens at least 20 each billion-dollar market applications heretofore considered economically or physically infeasible. Examples include full text search of both structured and unstructured without preprocessing for indexing, tagging or otherwise surrogating the real data; network intrusion detection; information assurance; fault detection in computer software; multi-sensor fusion; litigation support; military or business intelligence; healthcare record compliance; proteomics; DNA analysis; copyright infringement detection and others.

When implemented in semiconductor technology and used as an adjunct to a host microprocessor key GPSTP® capabilities are;

a) **Fast,** qualifies input bytes at approximately 1 Gb/sec because internal processing uses associative addressing and systolic propagation. A single chip will have an internal bandwidth of 2^{16} bits per clock cycle (as contrasted to 2^5 = 32 or 2^6 = 64 bits per cycle in current commercial computers).
b) **Constant speed** regardless of the complexity of the qualification criteria or number of relevant bytes in the input data. No combinatorial explosion of run time.
c) **Accurate,** a user can drive the incidences of false positives and false negatives to zero because qualification criteria can be arbitrarily extensive and complex.
d) **Price/performance** considering throughput/$ and watts/Gb/s, at least 10-fold better than can be achieved with current generation von Neumann class stored program computers.
e) **Scalable,** essentially unlimited in both capacity and deployed location because multiple chips can process the source data or respective subsets thereof.
f) **Massively parallel**, enabling qualification of source data with respect to arbitrarily large and complex criteria. A current implementation in semiconductor technology contains 65K recognizers each capable of discerning 256 distinct input byte values. Also a user-specifiable interconnection web among the recognizers supports 2^{65536} states that enable qualification of numerous relational patterns among the input bytes. The relevance of each input byte is judged not only on its value but also on its contextual relation to all previous bytes.
g) **Boolean and Set** operators are enabled.
h) **Ternary logic,** Yes, No and Doesn't Matter, among the recognizers allowing the

qualification process to gloss over 'noise' bytes in the input and 'censored' input. The quantity of bytes subject to the Doesn't Matter operator can be specified.

Like any hardware the GPSTP must be complemented by a suite of user-oriented software that facilitates management of the users' ontic space, prepares qualification criteria, configures the chip with the criteria, manages source data inspection sessions and dispatches of GPSTP 'finds' to application-specific sessions.

Although the GPSTP is new an earlier, simpler version of this kind of processor has been in operation in the intelligence community since the 1970's. It has proven superior to stored program computers in full text search and similar applications. It contains rather complicated latching, lattice logic which limited its functionality, extensibility and throughput as well as cost. In 1980 Curt Harris conceived a state-space transform that reduced the complex logic to a series of reads and writes in a memory. The notion of trading memory space for logic gates makes the GPSTP unique. Then the introduction of the GPSTP had to wait until the semiconductor industry could devise 1Gb chips.
Now is the time.

Figure 1 shows the variety of criteria the chip uses to qualify an input byte stream. The criteria are chosen by the user and processed by the control microprocessor to formulate a bit pattern that is loaded into the chip thereby configuring its recognizers and interconnection network. The control microprocessor then presents the input byte stream to the chip in a byte-sequential manner. Each input byte is examined not only for its value but also for the ways its arrival can change the relevance of subsequent bytes.

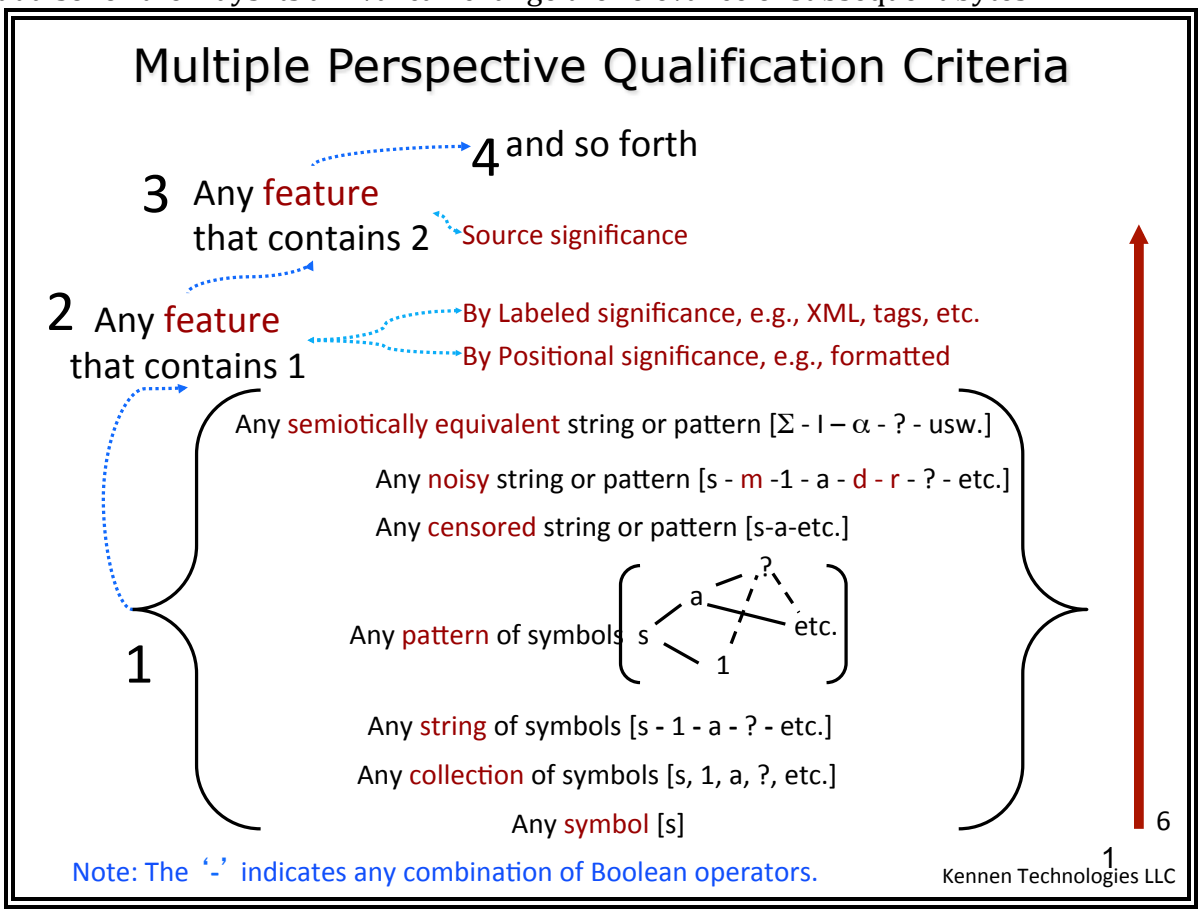

Figure 1. Qualification Criteria

Reading Figure 1 from the bottom up, a user can have the chip qualify

a) a specified symbol (single or multiple byte)
b) a collection of symbols both an instance of each kind or multiple instances of each kind.
c) an ordered set of two or more symbols where the sentential criteria is "follows" or "preceded by." Also called a string.
d) a pattern of symbols. Essentially knotted strings, a network topology of relevance.
e) a censored string or pattern. The '?' signifies a Doesn't Matter byte or bytes.
f) a noisy string or pattern. The '?' signifies a Doesn't Matter byte or bytes.
g) a semiotically equivalent string or pattern wherein an icon has an a priori byte value or is represented as an ordered set of bytes as discovered by the chip.
h) any feature in the source bytes as determined by a) through g) above can be further qualified by a labeled significance such as an XML tag or a positional significance in formatted input.
i) any feature classified by relevance of content and local context as in h) can be further classified as to source or other kinds of input context.
j) Any 'i)' can be further classified by user relevance such as degree of significance or urgency.

Figure 2 shows more detail regarding GPSTP capabilities and operations. The GPSTP can be applied not only to classical pattern matching/recognition wherein the expected pattern is specified to the chip but also to data qualification which only entails loading the chip with criteria for byte content and context from which it finds disparate content in the source data store or stream that have no *a-priori* relationships.

Figure 2 summarizes GPSTP® key capabilities, a) Detection cells, b) Cluster Recognizer, Threshold Cell, Threshold Aggregation, Cluster Response and Logical Aggregation.

The GPSTP can be implemented such that once configured the qualification criteria can be extended without stopping to reconfigure the chip. However, the chip is not reflexive which means the qualification criteria cannot be internally edited based on a finding in the input. Such edits must be done by the control computer that then reloads the GPSTP chip (est. microseconds).

In addition to using chips in parallel for greater capacity or for deploying chips to multiple source data locations, it is feasible to stack chips so that one layer can discern patterns in the findings of a lower layer that is processing the source data.

Current semiconductor technology limits one chip to 65K detection cells per device. Multiple chips can be run in parallel, particularly if it is important to locate the chips where the data is located rather than flowing massive amounts of data to a single location.

To start a session a GPSTP device is configured with a Reference Pattern by a User Host Device. Then source data is introduced in byte serial fashion and qualifying source data are found. For each find the GPSTP device flags the Host Device with a Recognized Object Response.

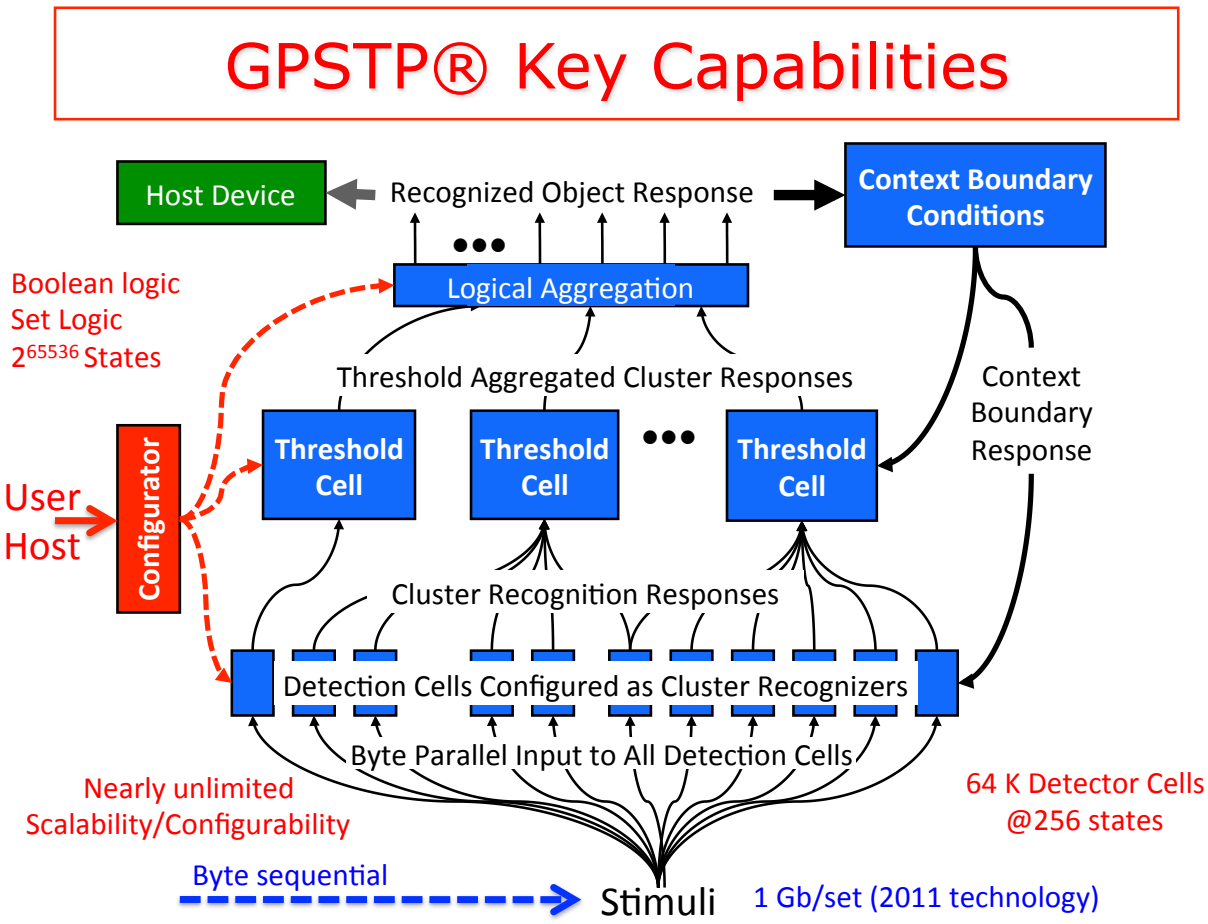

Figure 2. GPSTP Operational Concept

Reading bottom-up, the byte sequential source data is presented to all detection cells simultaneously. Each cell has three parts, an associative memory of 256 bits, a way of outputting satisfaction pointers to the systolic interconnection network, and a way of being activated or not by other Detection Cells via the network. The Cluster Recognizers can consist of one or more detection cells as specified by the qualification criteria.

If a source byte qualifies then all relevant Recognizers are activated for the next byte. As a next byte qualifies or not the Threshold Cells determine whether a string in the configured Reference Pattern is being satisfied. Also, this enables managing the duration of each Doesn't Matter situation in the Reference Pattern.

Multiple strings and sets of knotted strings may be configured by the reference pattern so that multiple users can be served with only one pass of the source data. The results are then fanned out to the right users.

<end>