Algorithms for Advanced Packet Classification with TCAMs

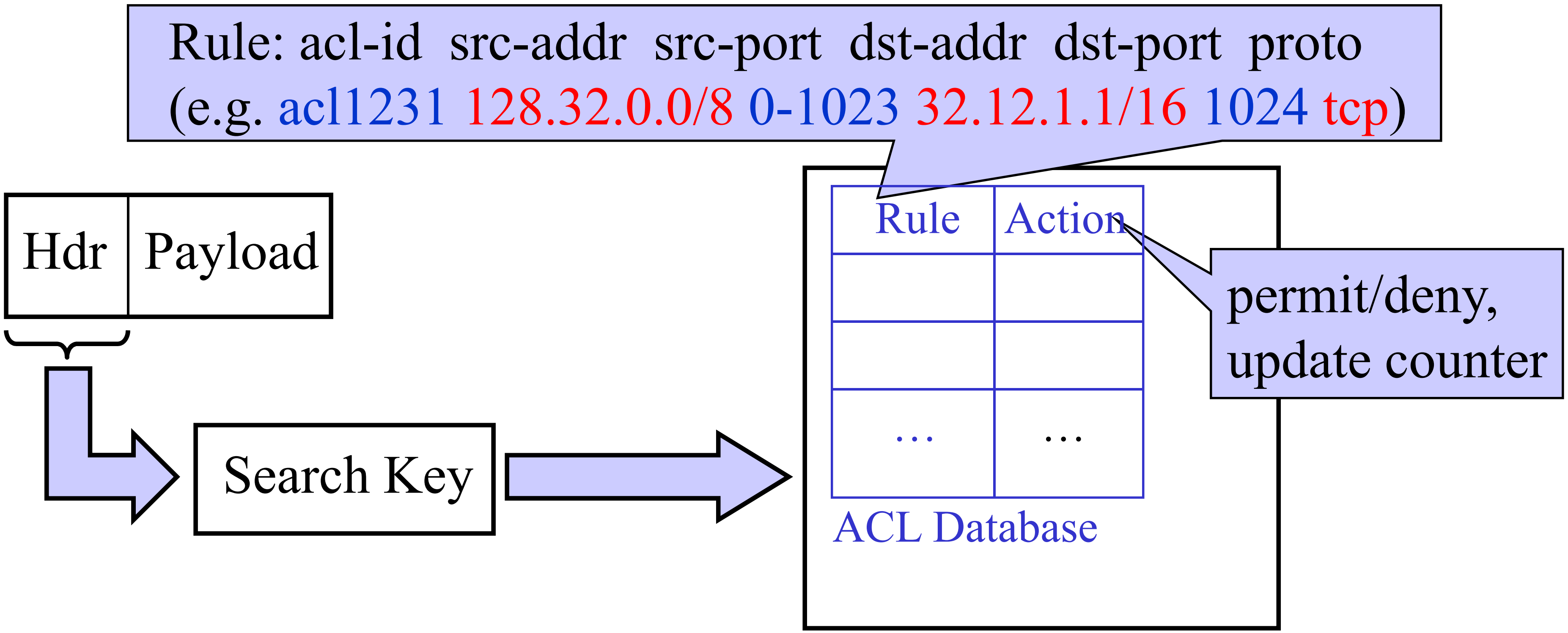
(sigcomm 2005)

Karthik Lakshminarayanan
UC Berkeley

Joint work with
Anand Rangarajan and Srinivasan Venkatachary
(Cypress Semiconductors Inc.)

Modified by Yaxuan Qi, for NSLab Seminar,
Tsinghua Univ. Beijing China
May 5, 2009

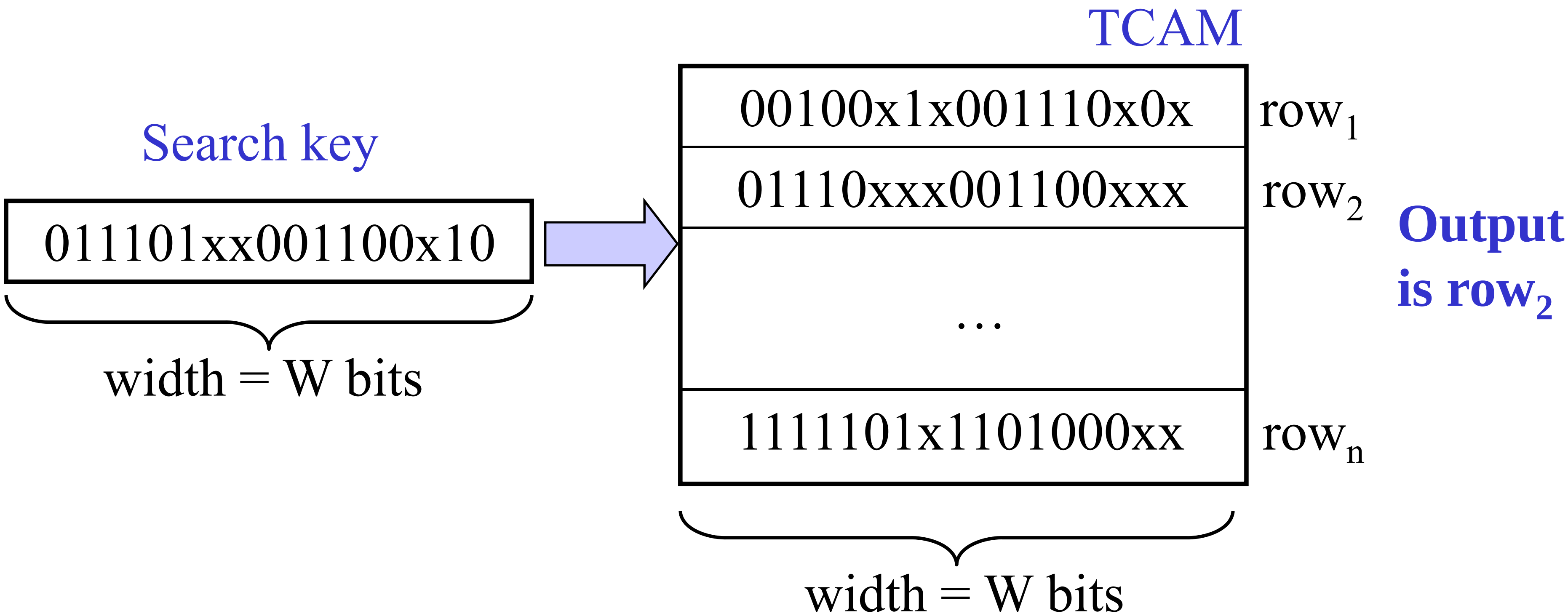
Packet Processing Environment



- Packet matches a set of rules based on the header
- Examples: routers, intrusion detection systems

Ternary Content Addressable Memory

- Memory device with fixed width arrays
- Each bit is 0, 1 or x (don't care)
- Search is performed against all entries in *parallel* and the *first result* is returned



TCAM: Benefits and Disadvantages

- Benefits:
  - Deterministic Search Throughput—O(1) search
- Disadvantages:
  - Cost
  - Power consumption
- Current TCAM usage:
  - 6 million TCAM devices deployed (by 2005)
  - Used in multi-gigabit systems that have O(10,000) rules
  - TCAMs can support a table of size 128K (18Mbits/144bits) ternary entries and 133 million (133M/15M=88Gbps 64B packets) searches per second for 144-bit keys

Range Representation Problem

- Representing prefixes in ternary is trivial
  - IP address prefixes present in rules
- Representing arbitrary ranges is not easy though
  - port fields might contain ranges
    - e.g. sPort [1024, 65536], dPort [6110, 6112]
  - intrusion detection may check packet length field
    - e.g. packet size [1, 254]
- Problem Statement
  - given a range R, find the minimum number of ternary entries to represent R

Why is efficient range representation an important problem?

Statistic	1998 database	2004 database
Total number of rules	41190	215183
With single range field	4236 (10.3%)	54352 (25.3%)
With single non-"$\geq$ 1024" range field	553 (1.3%)	25311 (11.8%)
With two range fields	0 (0%)	3225 (1.5%)
Unique ranges in first field	62	270
Unique ranges in second field	0	37

Number of unique ranges have increased over time

Earlier Approaches – I

Prefix expansion of ranges:

- express ranges as a union of prefixes
- have a separate TCAM entry for each prefix
- expansion: the number of entries a rules expands to

- Example: the range [3,12] over a 4-bit field would expand to:
  - 0011 (3), 01xx (4-7), 10xx (8-11) and 1100 (12)
- Worst-case expansion for a single W-bit field is 2W-2
  - example: [1,14] would expand to 0001, 001x, 01xx, 10xx, 110x, 1110
  - 16-bit port field expands to 30 entries
  - F W-bit fields is thus $(2W\text{-}2)^F$

Earlier Approaches – II

Database-dependent encoding:

– observation: TCAM array has some unused bits

– use these additional bits to encode commonly occurring ranges in the database

- TCAMs with IP ACLs have ~ 36 extra bits
  - 144-bit wide TCAMs
  - 104-bits + 4-bits for IP ACL rules

Earlier Approaches – II

Database-dependent encoding:

– observation: TCAM array has some unused bits

– use these additional bits to encode commonly occurring ranges in the database

- Example:

Address	Port	…		
12.123.0.0/16	20-24	…	⟶	**Set extra bit to 1**
32.12.13.0/24	1024-	…	⟶	**Set extra bit to x**
128.0.0.0/8	20-24	…	⟶	**Set extra bit to 1**

If search key falls in 20-24, set extra bit to 1, else set it to 0

Earlier Approaches – II

Database-dependent encoding:

- observation: TCAM array has some unused bits
- use these additional bits to encode commonly occurring ranges in the database

- Disadvantages:
  - extra bits is limited
  - number of unique ranges is increasing
  - incremental update is hard
  - …
  - all due to: database dependency

Database-Independent Range Pre-Encoding

- Key insight: use additional bits in a database independent way
  - wider representation of ranges
  - reduce expansion in the worst-case

Database-Independent Range Pre-Encoding

- Fence encoding (W bits):
  - total of 2^W-1 bits
  - encoding of *i* has *i* ones preceded by 2^W-*i*-1 zeros
  - e.g. W=3, f(0) = 0000000, f([1, 3]) = 0000xx1
- With 2^W-1 bits, fence encoding achieves an expansion of 1

Range	Encoding
$= i$	$0^{2^k-i-1}1^i$
$\geq i$	$x^{2^k-i-1}1^i$
$< i$	$0^{2^k-i}x^{i-1}$
$[i, j]$	$0^{2^k-1-j}x^{j-i}1^i$

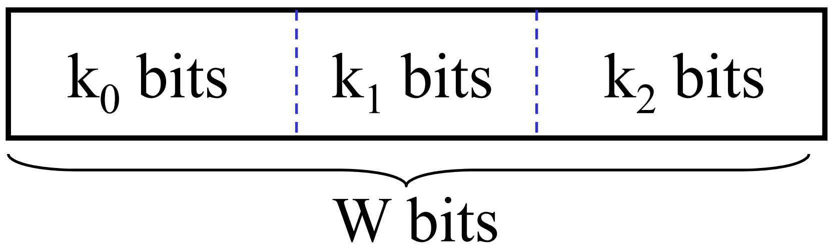
Theorem: For achieving a worst-case row expansion of 1 for a W-bit range, 2^W-1 bits are necessary

DIRPE: Using the Available Extra Bits

- Two extremes:
  - no extra bits → worst case expansion is $2W-2$
  - 2^W-W-1 extra bits → worst case expansion is 1
- Is there something in between?
  - appropriate worst-case based on number of extra bits available

Database-Independent Range Pre-Encoding

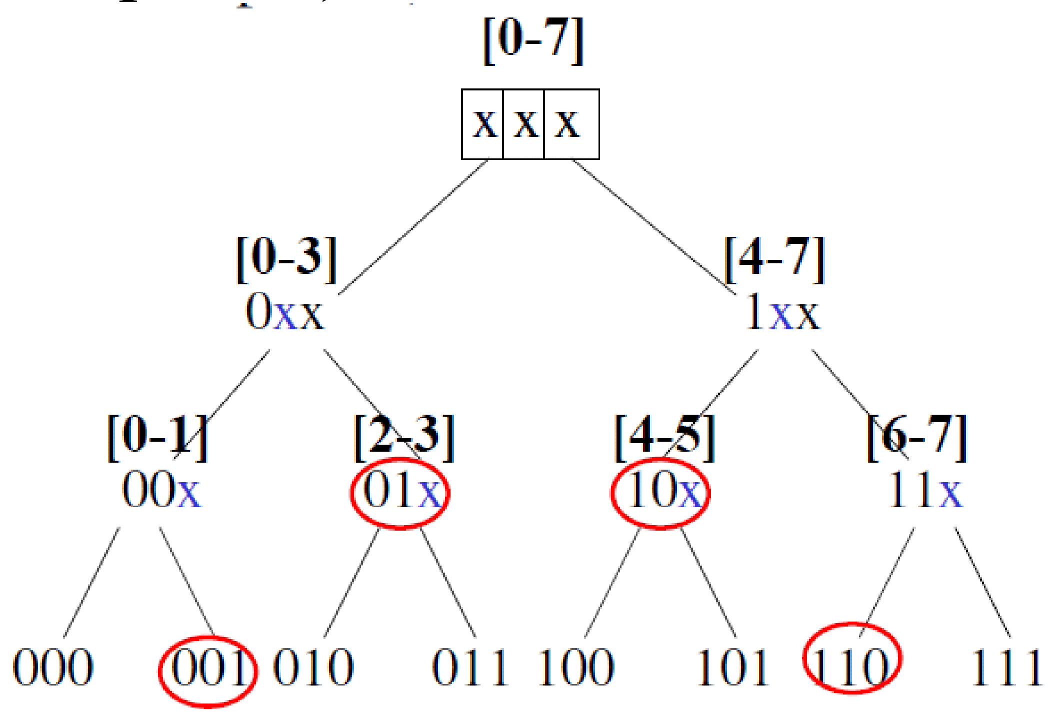
- Procedure:
  - split W-bit field into multiple *chunks*
  - encode each chunk using fence encoding
  - “combine” the chunks to form ternary entries



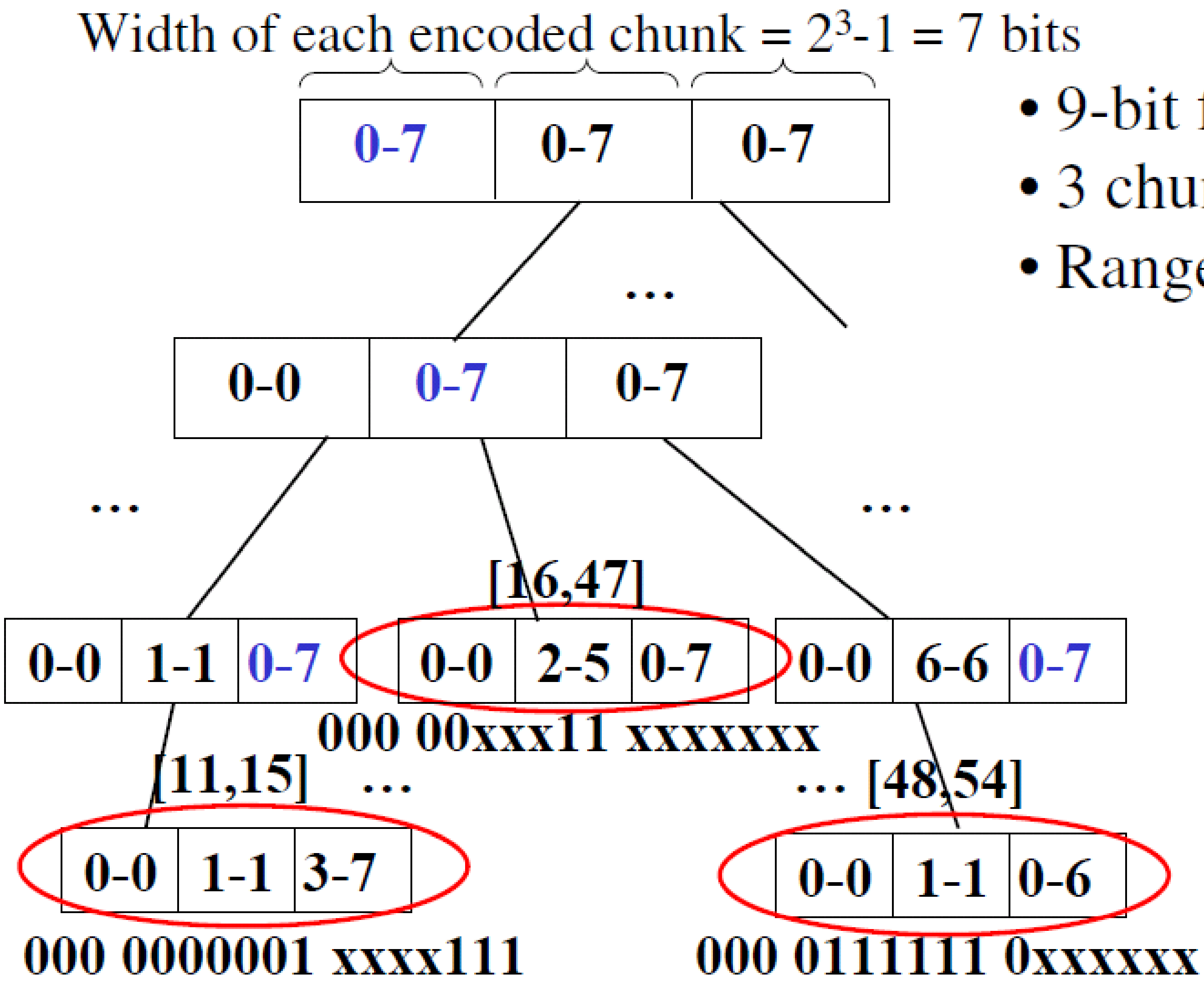
Combining chunks: analogous to multi-bit tries

Unibit view of DIRPE (Prefix expansion)

- W=3 divided into 3 one-bit chunks
- R=[1,6]—prefixes = {001,01x,10x,110}
- Each level can contribute to at most 2 prefixes (but the top level)



Multi-bit view of DIRPE



- 9-bit field (W=9)
- 3 chunks, 3 bits wide
- Range = [11,54]
  = [013, 066]

Worst case expansion = 2W/k – 1

Number of extra bits needed = $(2^k-1)W/k - W$

Comparison of Expansion

Extra bits	DIRPE	Region-based Range Encoding
0	30	30
8	15	30
18	11	16
27	9	14
44	7	12

Worst-case expansion

Extra bits	DIRPE	Region-based Range Encoding
0	2.69	2.69
8	2.08	2.33
18	1.79	2.17
36	1.57	1.58

Real-life expansion

DIRPE + DB-dependent → Net expansion was 1.12

Metric	Prefix Expansion	Region-based Encoding (with *r* regions)	DIRPE (with *k*-bit chunks)	DIRPE + Region-based
Extra bits	0	$F(\log_2 r + \frac{2n-1}{r})$	$F(\frac{W(2^k-1)}{k} - W)$	$F(\frac{(2^k-1)\log_2 r}{k} + \frac{2n-1}{r})$
Worst-case capacity degradation	$(2W-2)^F$	$(2\log_2 r)^F$	$(\frac{2W}{k} - 1)^F$	$(\frac{2\log_2 r}{k})^F$
Cost of an incremental update	$O(W^F)$	$O(N)$	$O((\frac{W}{k})^F)$	$O(N)$
Overhead on the packet processor	None	Pre-computed table of size: $O((\log_2 r + \frac{2n-1}{r}) F.2^W)$ (*or*) $O(nF)$ comparators of width W bits	$O(\frac{W.2^k}{k})$ logic gates	Both pieces of logic from previous two columns

DIRPE: Summary

↑ Database independent

↑ Scales well for large databases

↑ Good incremental update properties

↓ Additional bits needed

↓ Small logic needed for modifying search key

Related Work I

- Range-to-prefix conversion
  - Represent a range by a set of prefixes, each of which can be stored by a single TCAM entry. *(V. Srinivasan, G. Varghese, S. Suri, and M. Waldvogel, "Fast and scalable layer four switching," in ACM SIGCOMM, Sep. 1998, pp. 191–202.)*
  - The worst-case expansion ratio is 2W−2, in a single dimension.
  - A single rule can generate up to 900 prefixes (only for the two port fields).
  - prefix expansion may increase the number of required TCAM entries by a factor of more than 6.
- Direct hardware solution
  - Extended TCAMs, implements range matching directly in hardware. *(E. Spitznagel, D. Taylor, and J. Turner, "Packet classification using extended TCAMs," in ICNP, 2003.)*
  - Reducing power consumption by over 90% relative to standard TCAM
  - Will not be accomplished in the near future

Related Work II

- Database-dependent range encoding algorithms
  - Encoding is a function of the distribution of ranges in the database
  - Basic idea: a single extra bit is assigned to each selected range *r* in order to avoid the need to represent *r* by prefix expansion
    - the number of unique ranges in today' s databases is ~300
    - we have ~30 extra bits…
  - Region Partition: split a range into multiple sub-ranges. Each such sub-range is encoded by two numbers: the region number into which it falls, and the sub-range number within that region. *(H. Liu, "Efficient mapping of range classifier into ternary-cam," in Hot Interconnects, 2002.)*
  - Dynamic Range Encoding (DRES): a greedy algorithm that assigns extra bits to the ranges with highest prefix expansion. *(H. Che, Z. Wang, K. Zheng, and B. Liu, "Dres: Dynamic range encoding scheme for tcam coprocessors," IEEE Transaction on Computers, vol. 57, no. 6, 2008.)*
  - Layered Interval Coding (LIC): a more efficient representations based on the observation that, sets of disjoint ranges may be encoded much more efficiently than sets of overlapping ranges. *(Anat Bremler-Barr, David Hay, Danny Hendler, Beer-Sheva and Boris Farber, " Layered interval codes for tcam-based classification, INFOCOM 2009.)*

Related Work III

- Database-independent range encoding algorithms
  - Encoding of a specific range does not change across different databases.
  - Fence coding: just presented. *(K. Lakshminarayanan, A. Rangarajan, and S. Venkatachary, "Algorithms for advanced packet classification with ternary CAMs," in ACM SIGCOMM, 2005.)*
  - Grey coding: based on the observation that small ranges, which occur frequently in real-world databases, are encoded more efficiently. *(A. Bremler-Barr and D. Hendler, "Space-efficient tcam-based classification using gray coding," in IEEE INFOCOM, 2007, pp. 1388–1396.)*

Thanks!

Q & A