

PARA
Research
DISE
Laboratory



uOttawa

Presenter: Yunfeng Gu

Supervisor: Azzedine Boukerche

PARADISE Research Laboratory

University of Ottawa, Canada

Supporting Multi-dimensional Range Query in the P2P Network

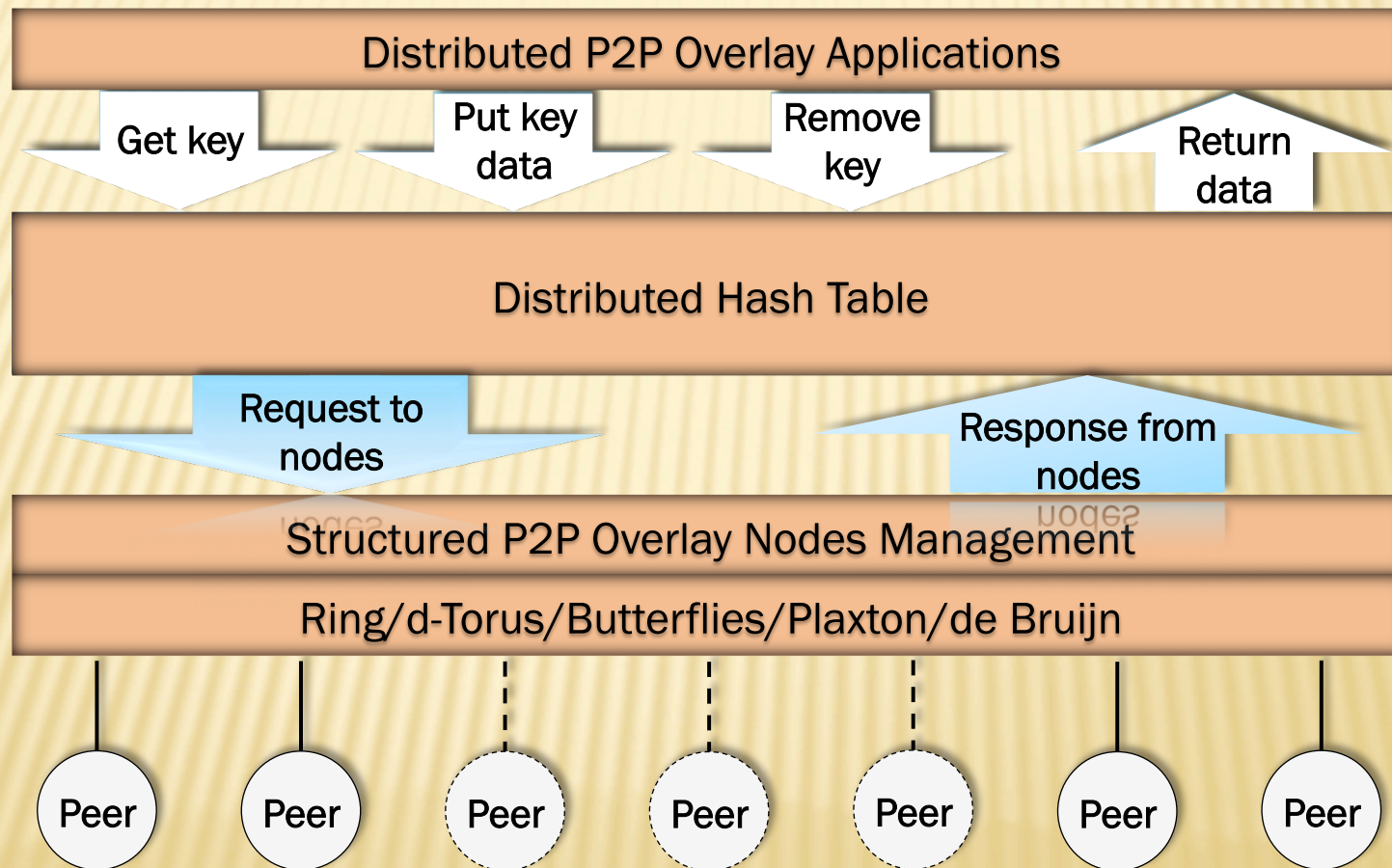
ERROR-RESILIENT ROUTING IN HD TREE

PEER-TO-PEER OVERLAY NETWORKS

✕ P2P Applications

- + Distributed directory systems
- + P2P domain name services
- + Network file systems
- + Massively parallel systems
- + Distributed e-mail systems
- + Massive-scale fault-tolerant system
- + ...

PEER-TO-PEER OVERLAY NETWORKS

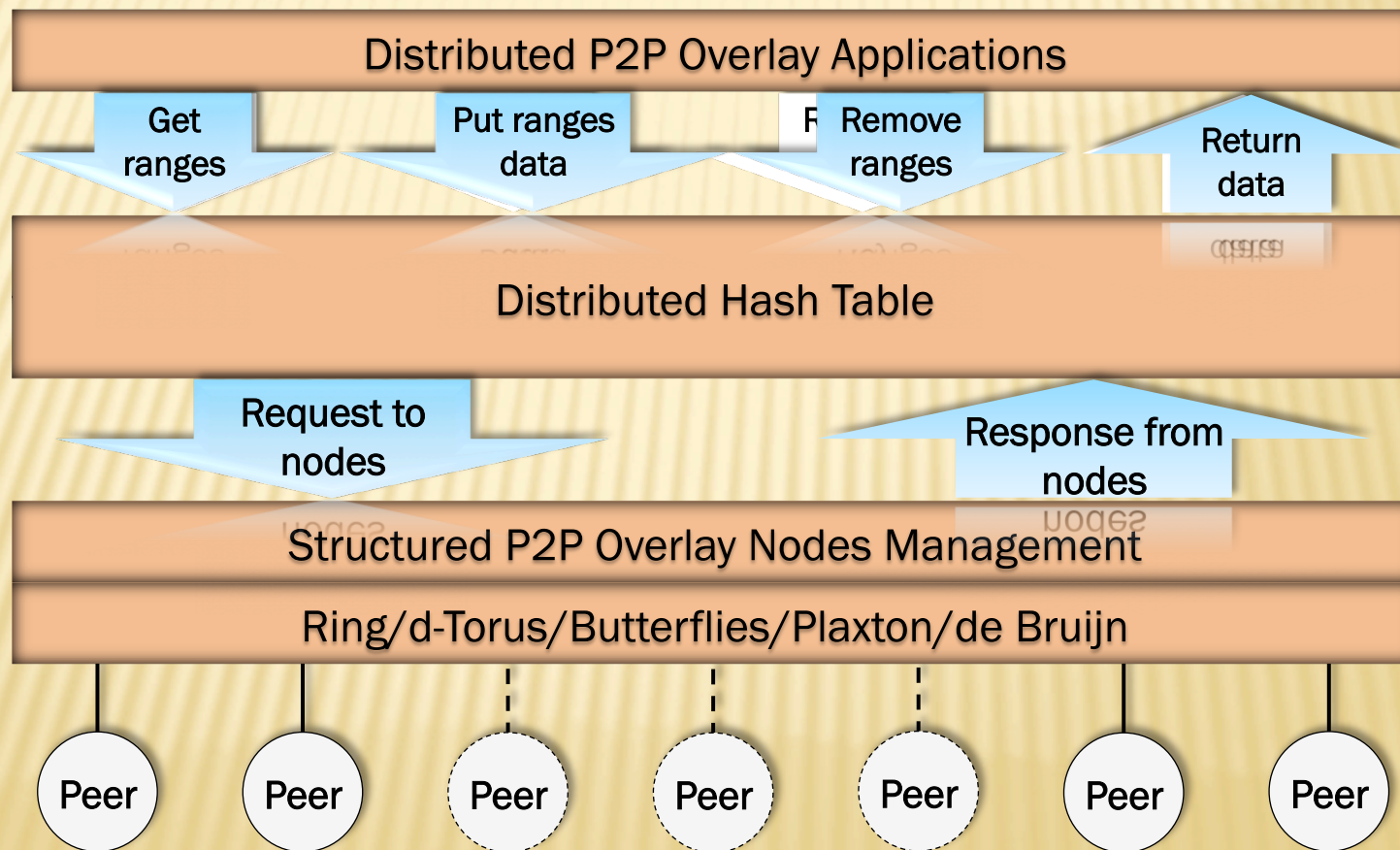


MULTI-DIMENSIONAL RANGE QUERIES

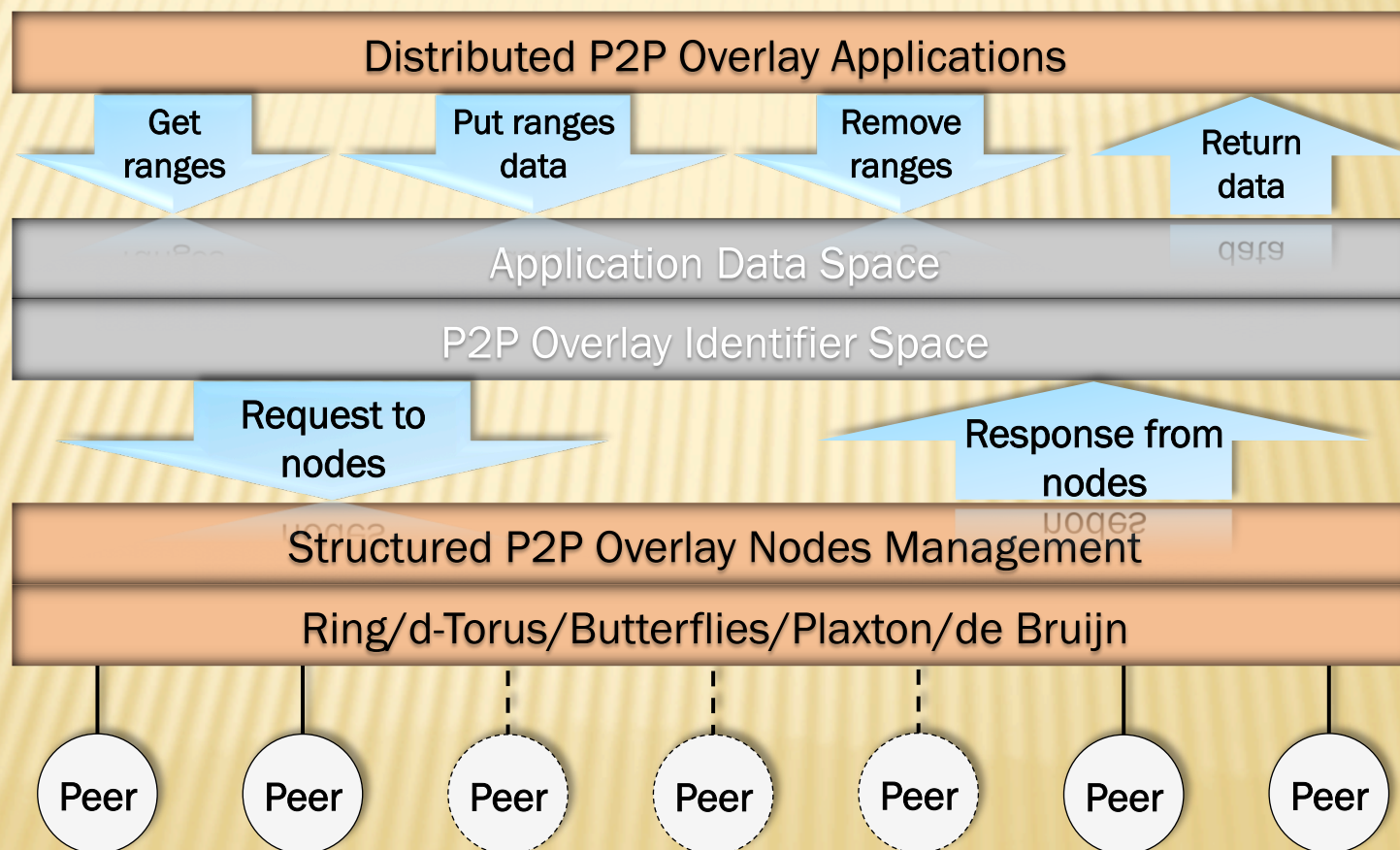
✕ Applications

- + Multi-player games
- + Grid computing
- + Publish/Subscribe systems
- + Group communications
- + Name services
- + P2P data sharing
- + Global storage
- + ...

MULTI-DIMENSIONAL RANGE QUERIES

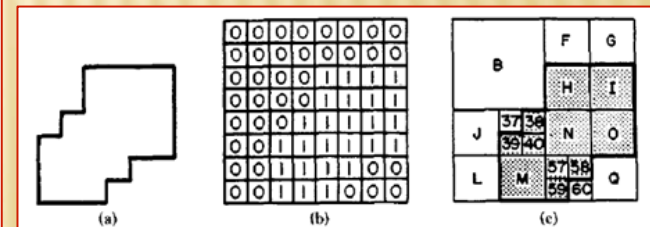
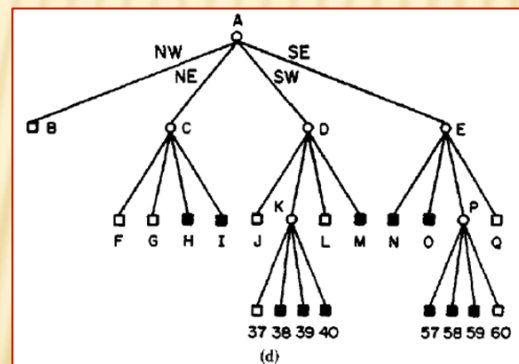
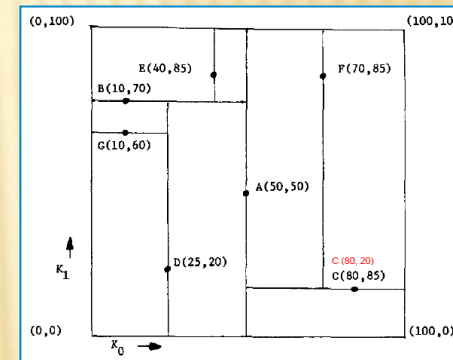
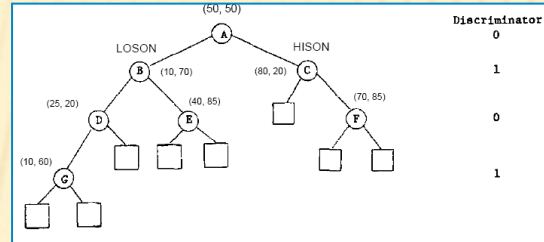


MULTI-DIMENSIONAL RANGE QUERIES



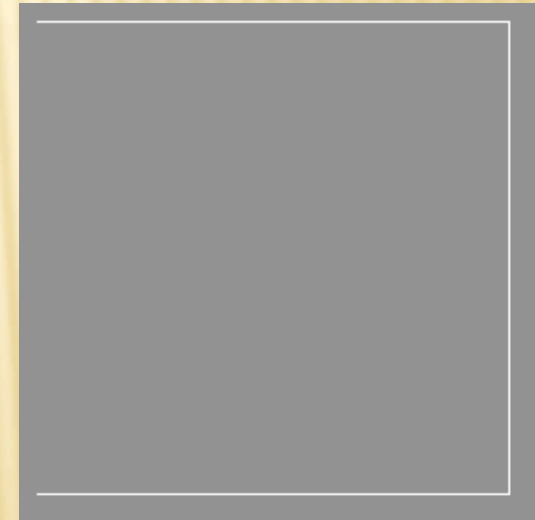
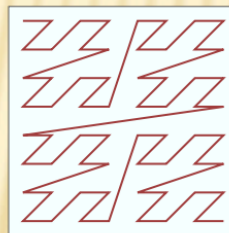
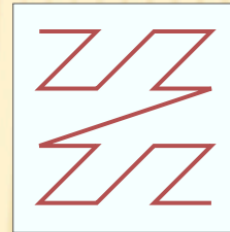
DATA SPACE PARTITIONING

- ✗ Goal
 - + Better preserve data localities
- ✗ Schemes
 - + K-d tree_[1]
 - + Quad tree_[2, 3]



DATA SPACE PARTITIONING

- ✖ Goal
 - + Better preserve data localities
- ✖ Schemes
 - + K-d tree_[1]
 - + Quad tree_[2, 3]
 - + Z-order SFC_[4, 6]
 - + Hilbert SFC_[5]

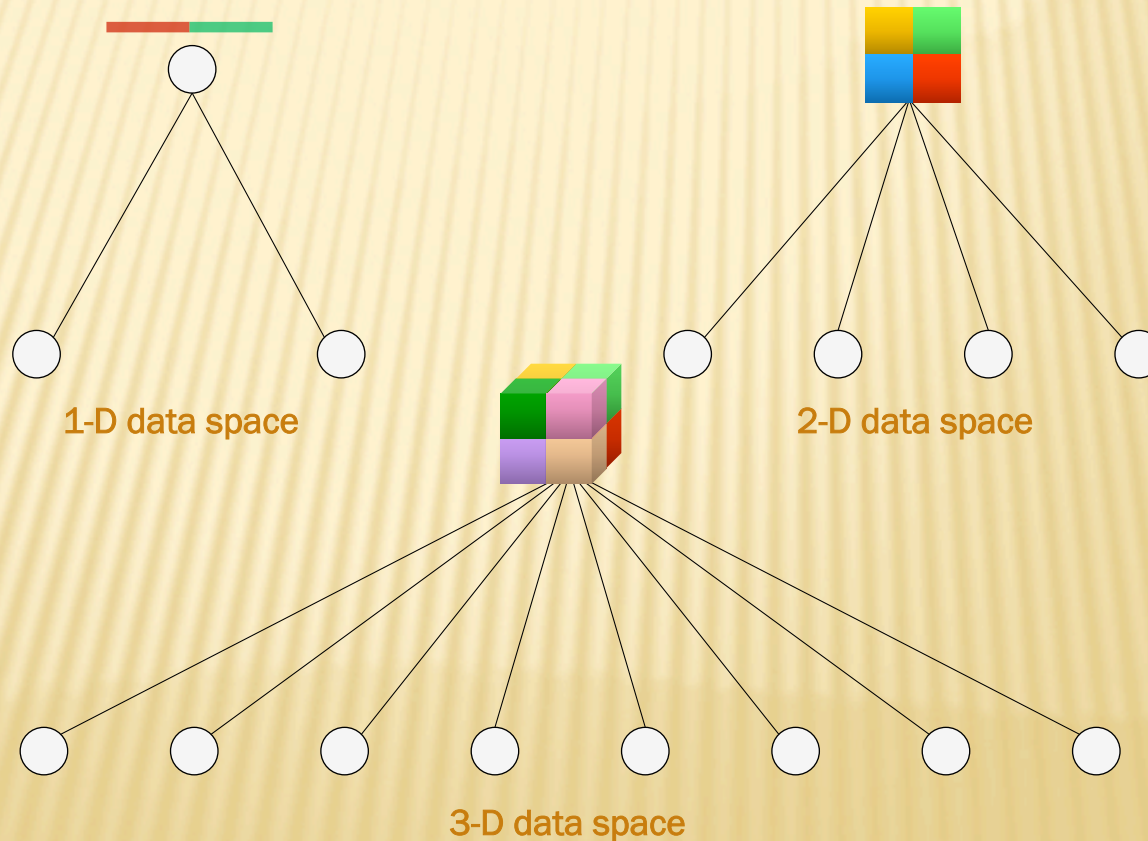


Copy from Wiki

DATA SPACE PARTITIONING

✖ Observations 1

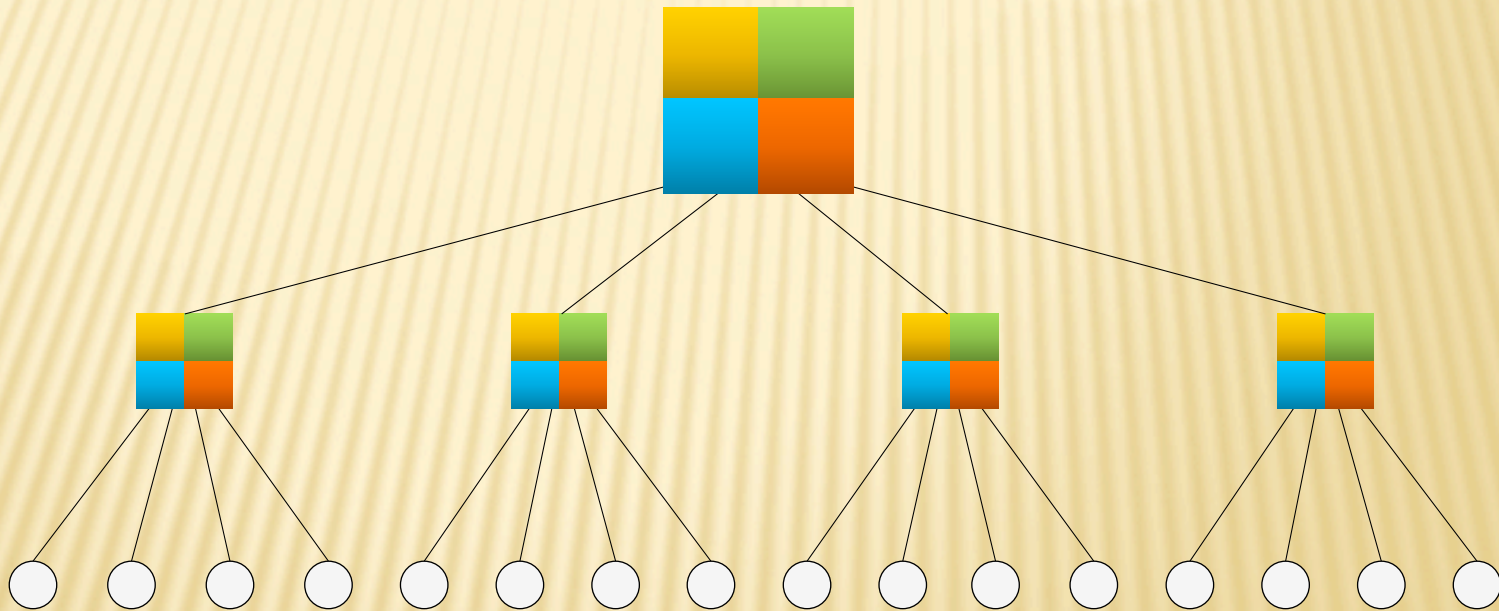
- + Data localities expand exponentially as the dimensionality of data space increases



DATA SPACE PARTITIONING

✗ Observations 2

- + Data localities extend exponentially as the order of the recursive decomposition increases



DATA SPACE MAPPING

- ✖ Challenge
 - + How to accommodate and maintain data localities with an exponentially expanding and extending rate at P2P layer
- ✖ Goal
 - + MDRQ can be done at a reasonable routing cost

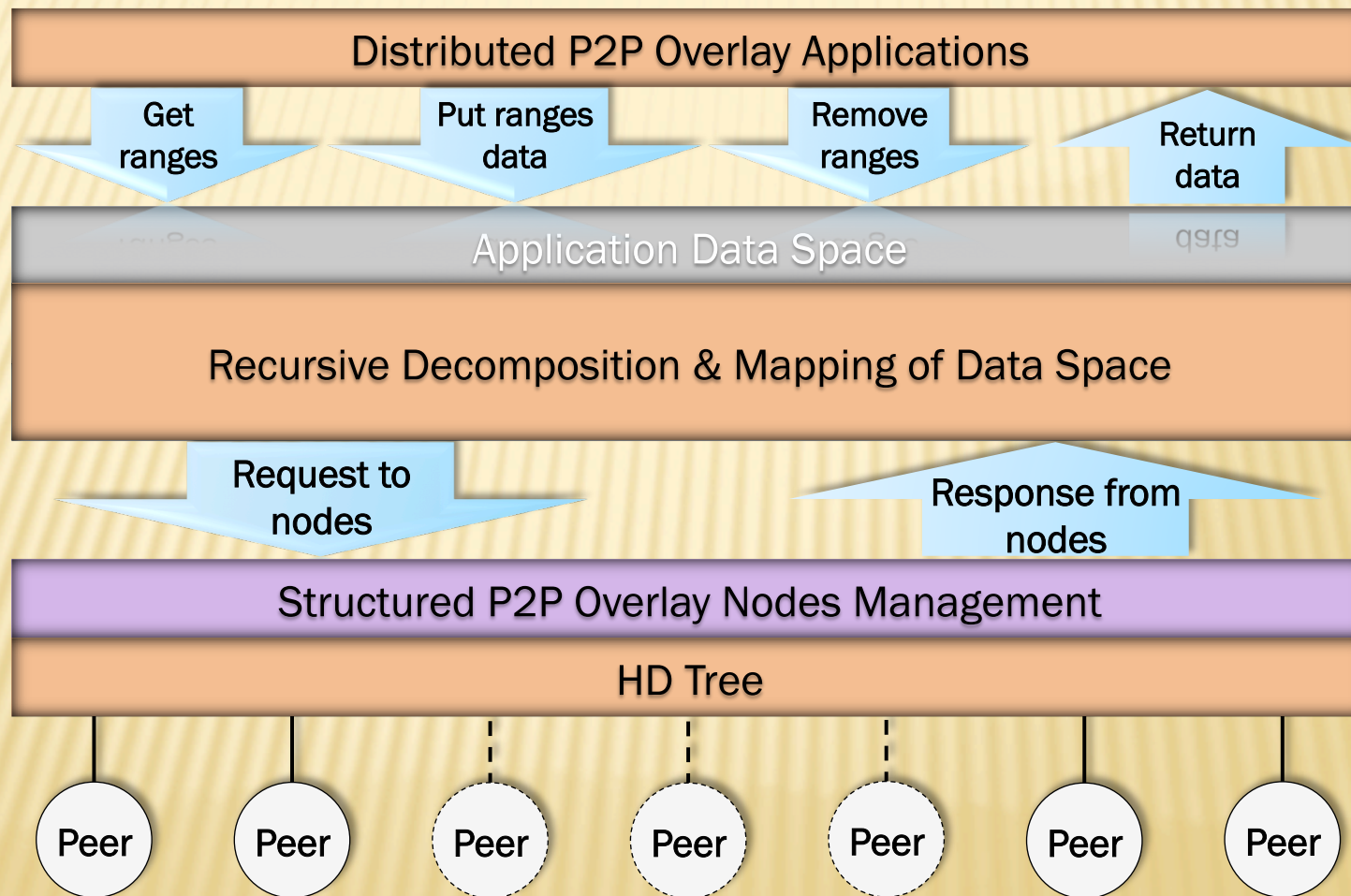
DATA SPACE MAPPING

P2P Systems	Partitioning	Underlying architecture	Direct mapping
SCARP _[10]	Z-order/Hilbert SFC	Skip graphs _[7]	No
MURK _[10]	K-d tree	d-torus (CAN) _[8]	No
ZNet _[11]	Z-order/Quad tree	Skip graphs	No
Skipindex _[12]	K-d tree	Skip graphs	No
Squid _[13]	Hilbert SFC	Ring (Chord) _[9]	No
SONAR _[14]	K-d tree	d-torus (CAN)	No

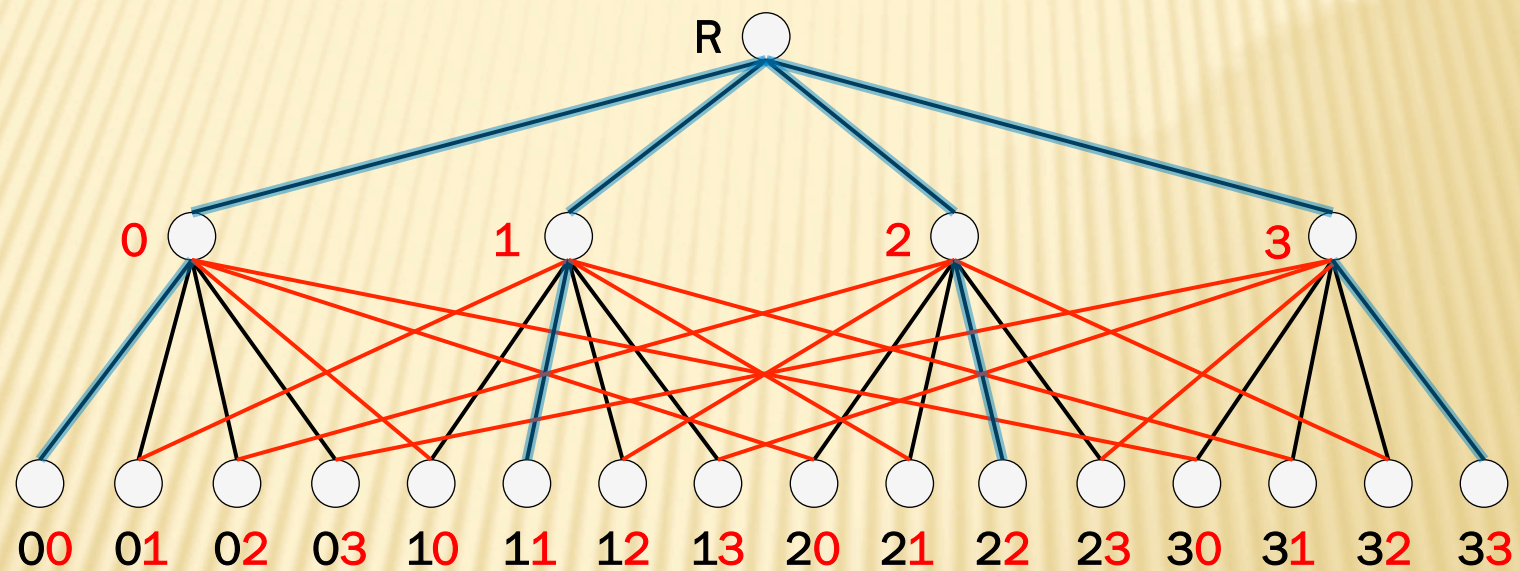
REMARKS

- ✖ MDRQ are underlying topology sensitive
- ✖ Existing P2P structures show no consistency with recursive process employed for data space decomposition at higher layer
- ✖ The inconsistency makes
 - + considerable maintenance and routing cost in existing P2P systems
 - + a simple, reliable, and scalable P2P solution for MDRQ very difficult

HIERARCHICALLY DISTRIBUTED TREE

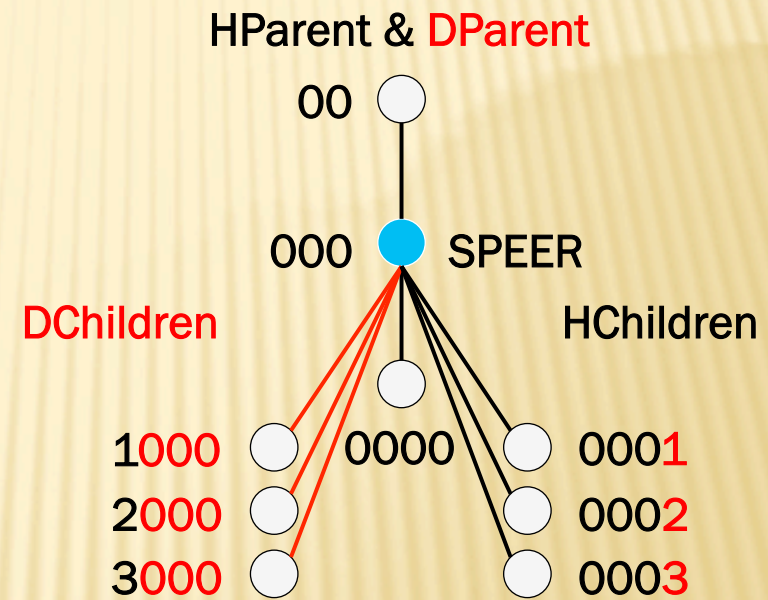
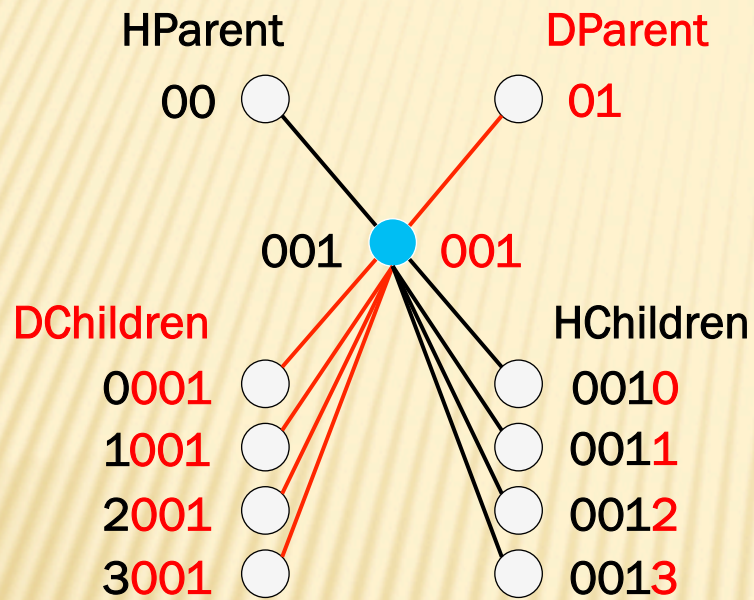


HD TREE

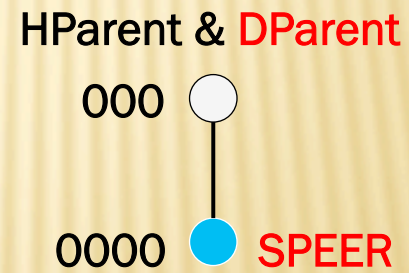
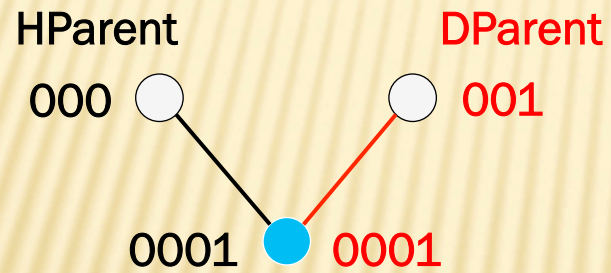
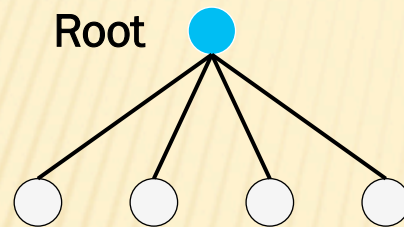


4-ary HD tree

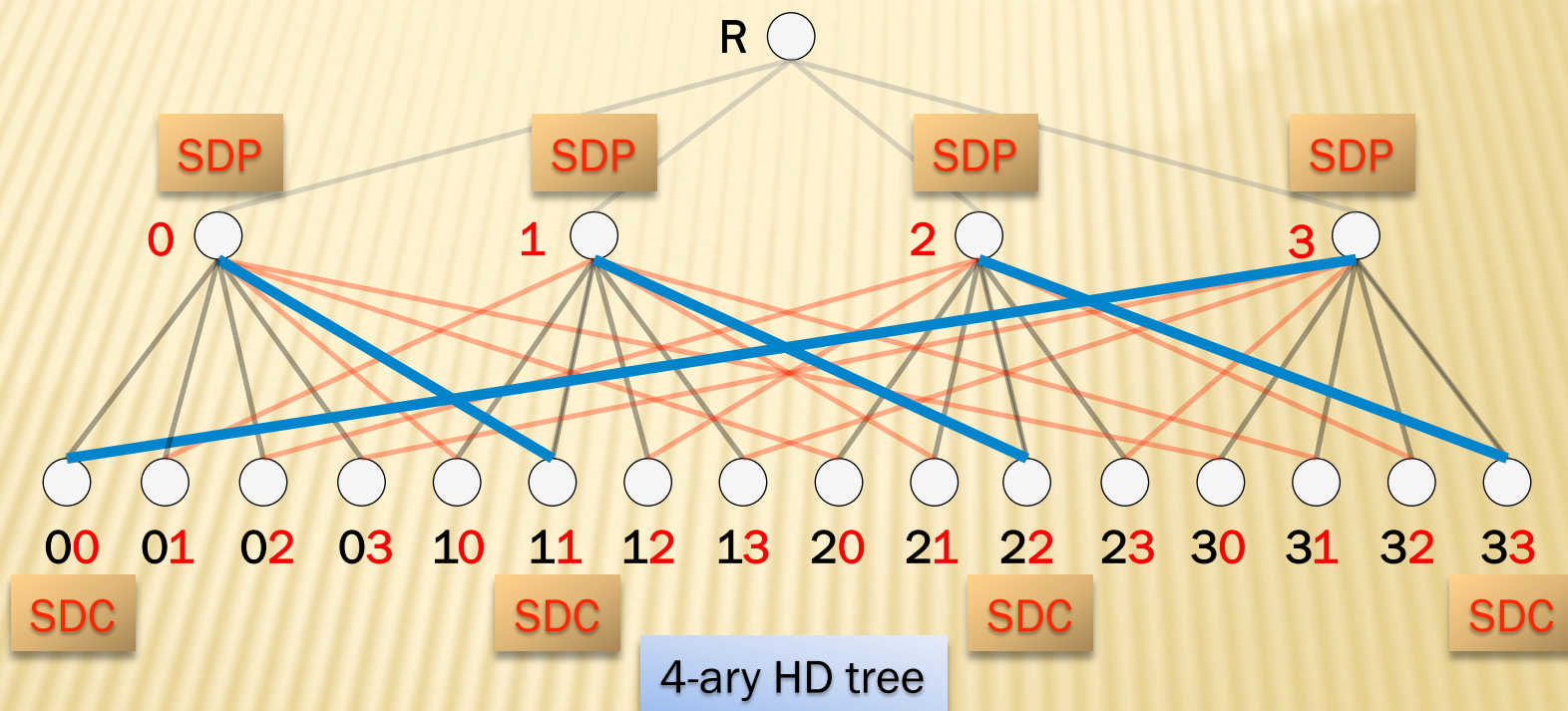
HD TREE



HD TREE



CONNECTING SPEERS IN HD TREE



HD TREE

Data structure	Routing table size (Neighbors)	Total links
HD Tree	$\leq 2(k + 1)$	$2(n - 1) - k \times h$
Tree	$\leq k + 1$	$n - 1$
Chord	$\log(n)$	?
CAN	$2d$	?
Skip graphs	$O(m \cdot \log(m))$	?

k: *k*-ary tree

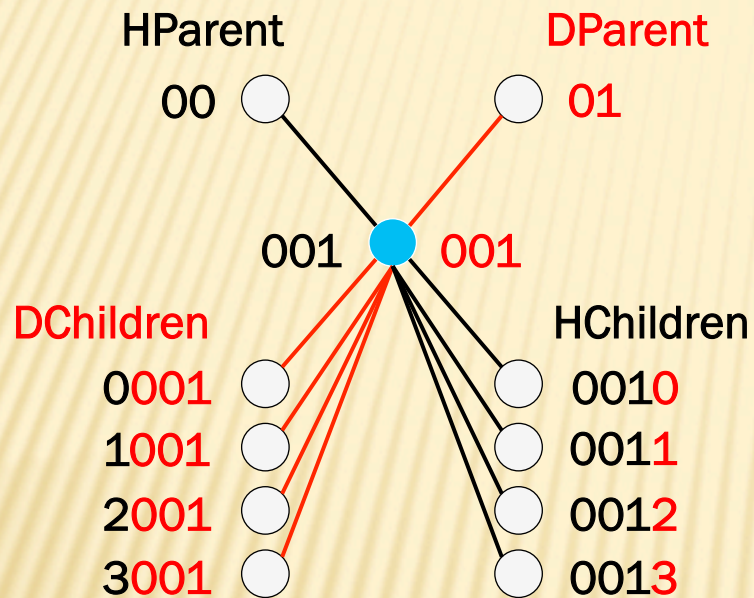
h: height of tree

n: total nodes in the system

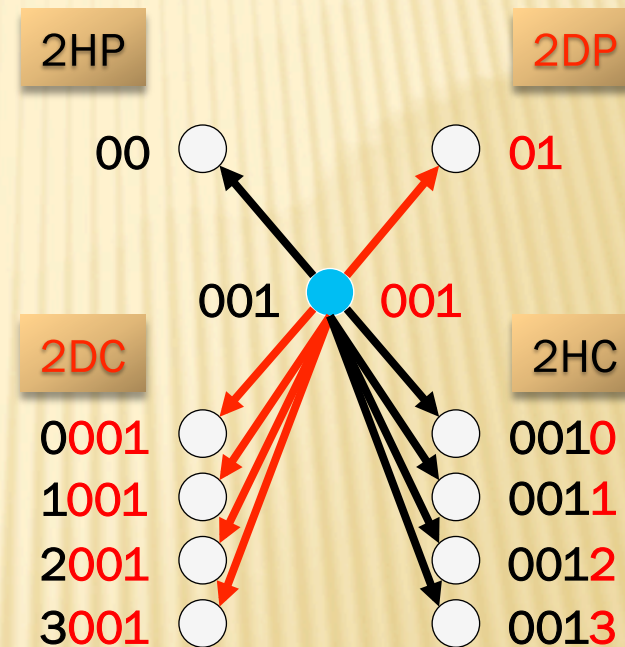
d: d-torus, number of dimensions

m: total number of data elements

HIERARCHICAL OPERATIONS



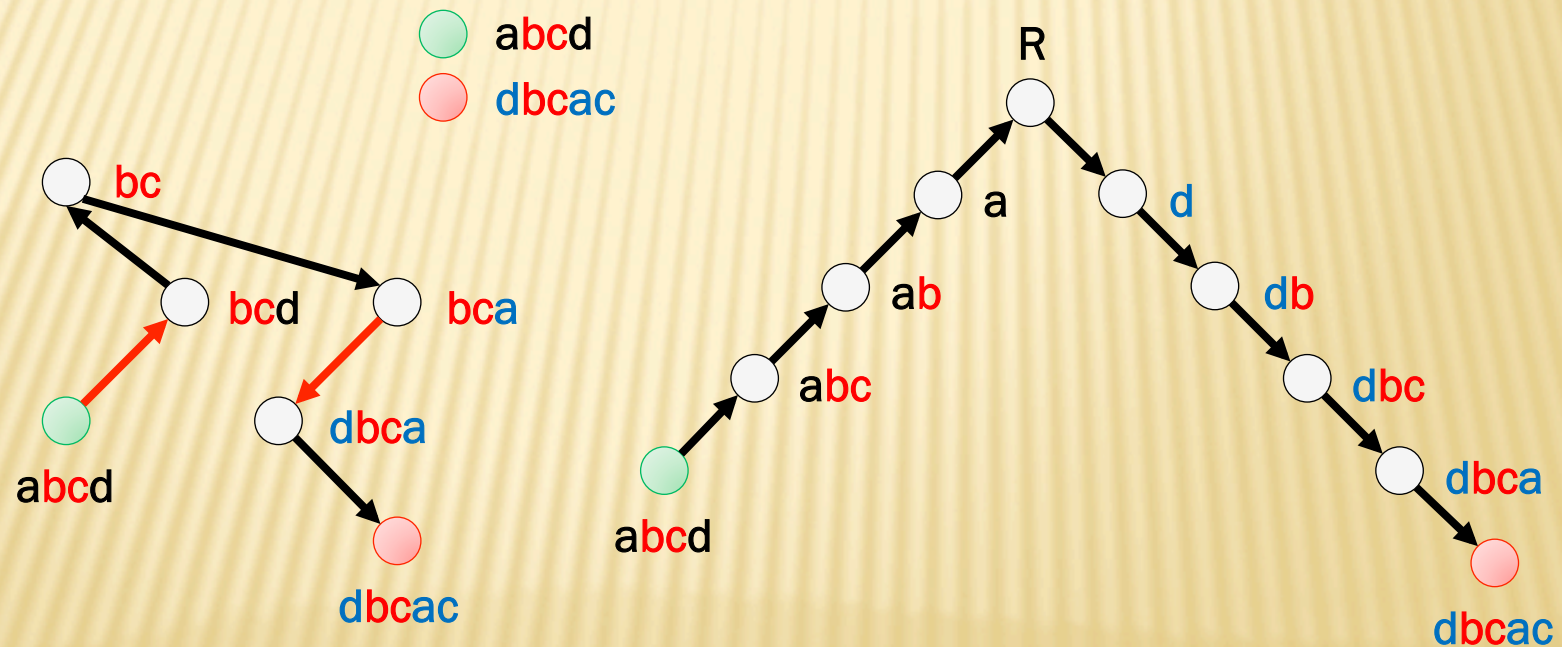
- ✗ A node in k-ary HD tree can have at most k HChildren and k Dchildren
- ✗ But only 1 HParent and 1 Dparent



- ✗ 2HP and 2DP has only one option
- ✗ 2HC and 2DC has at most k options

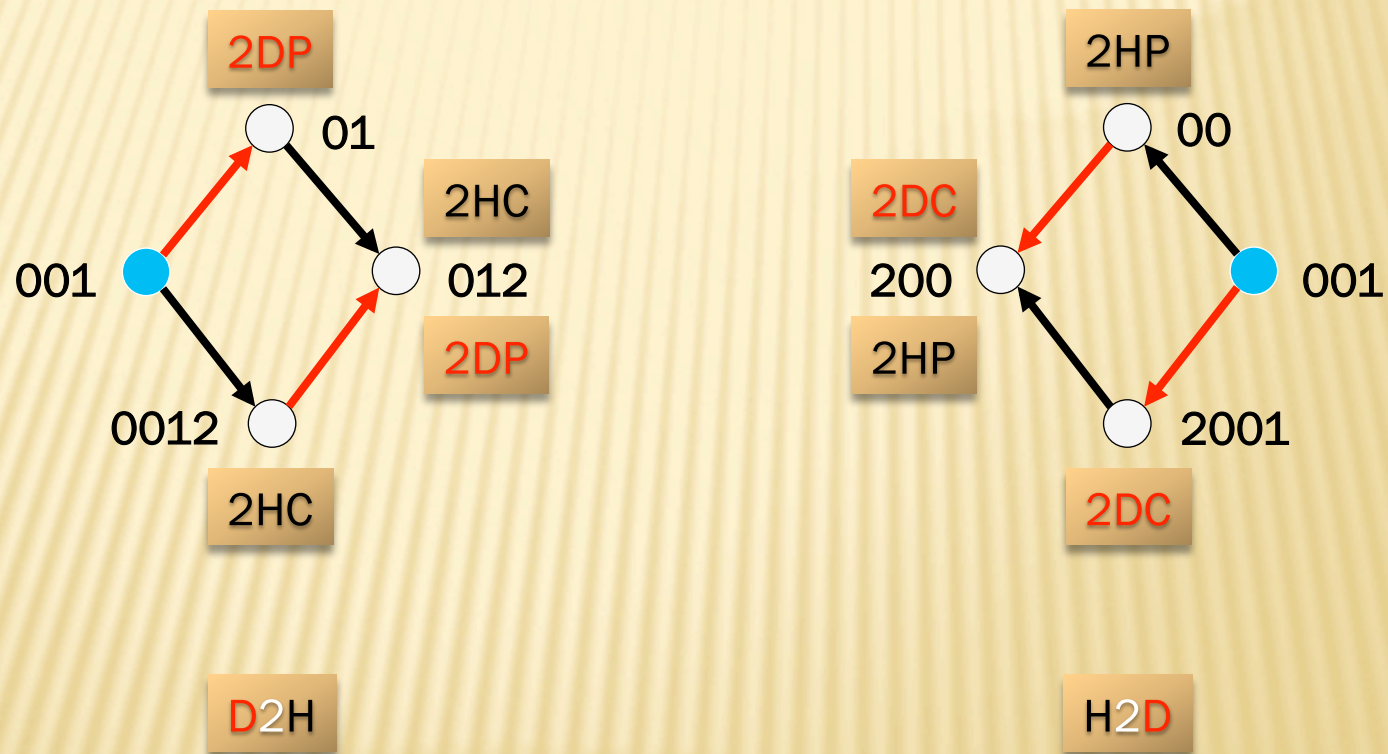
HIERARCHICAL ROUTING

- ✗ Hierarchical Routing (HR) consists of
 - + a series of 2HP or 2DP operations, or
 - + a series of 2HC or 2DC operations



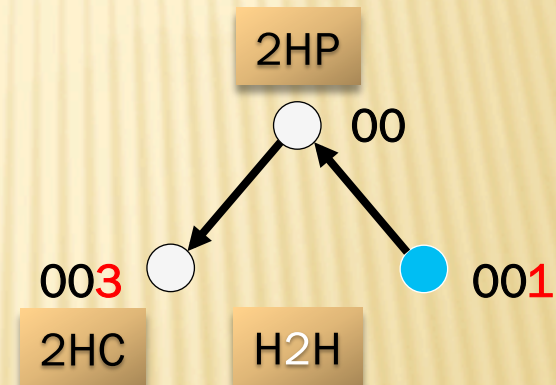
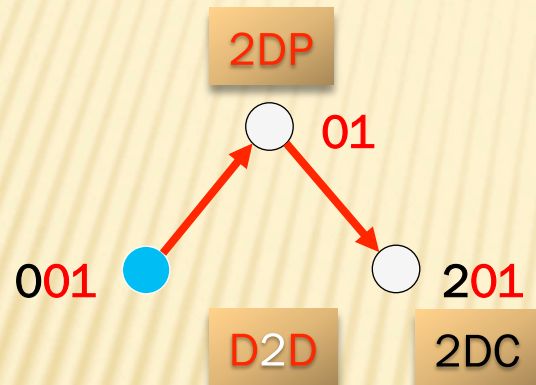
DISTRIBUTED OPERATIONS

- ✗ A distributed operation consists of two hierarchical operations



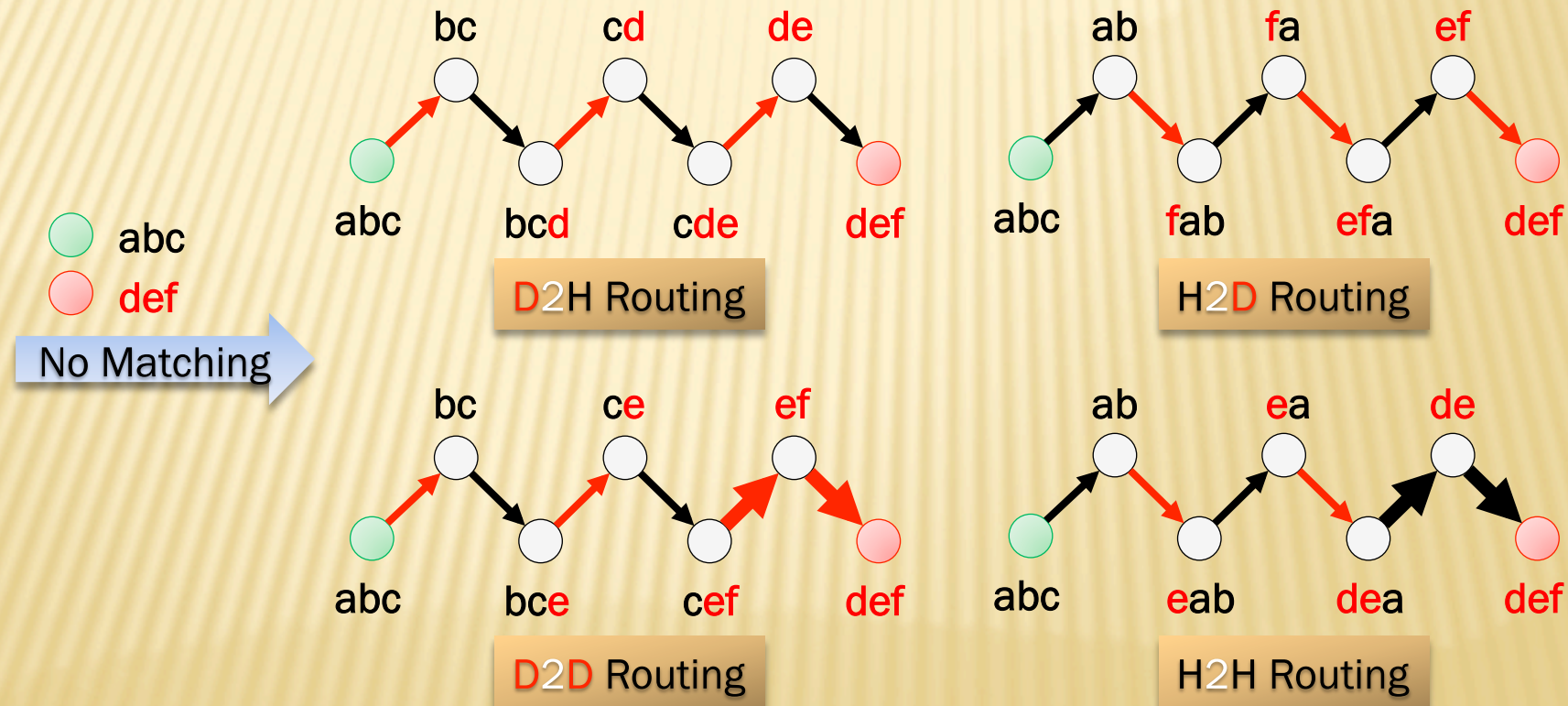
DISTRIBUTED OPERATIONS

- ✗ A distributed operation consists of two hierarchical operations



DISTRIBUTED ROUTING

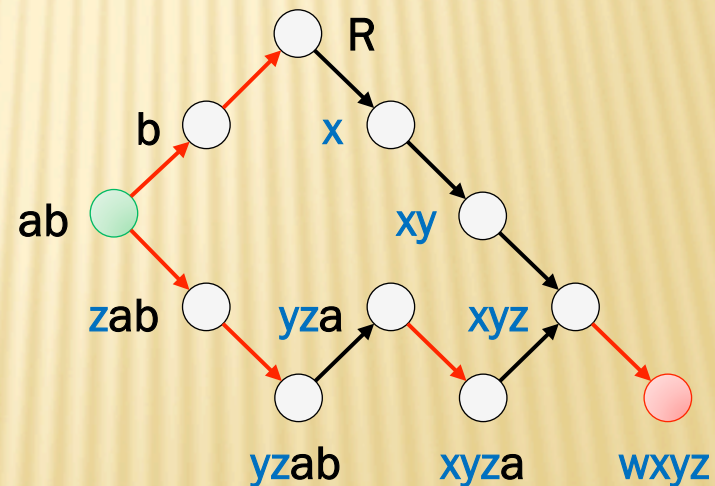
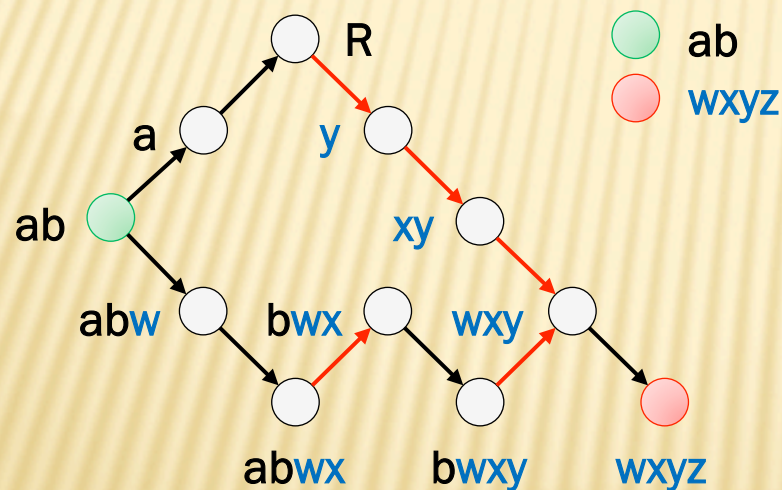
- ✗ Distributed routing (DR) is a series of distributed operations
- ✗ DR can reach the DST node in $(2 \times d)$ hops time at worst



Note: d = depth of node in HD tree = length of the code

DISTRIBUTED ROUTING

- ✗ DROCR is a DR oriented combined routing algorithm
- ✗ DROCR can reach any node in $(d_{SRC} + d_{DST} - 2 \times m)$ hops time
- ✗ Many routing strategies can be built over different routing options in HD Tree



Note: d_{SRC} , d_{DST} = depth of SRC/DST node in HD tree = length of the SRC/DST code
 m = length of the max. match between SRC and DST code

ROUTING PERFORMANCE SUMMARY

		Tree	D2H	D2D	H2D	H2H	HR	DROCR
DR	D2H	>	=	<	=	<	<	<
	D2D	>	>	=	>	=	<	<
	H2D	>	=	<	=	<	<	<
	H2H	>	>	=	>	=	<	<
HR		>	>	>	>	>	=	<
DROCR	HR + DR	>	>	>	>	>	>	=

>: performs better <: performs worse =: equivalent

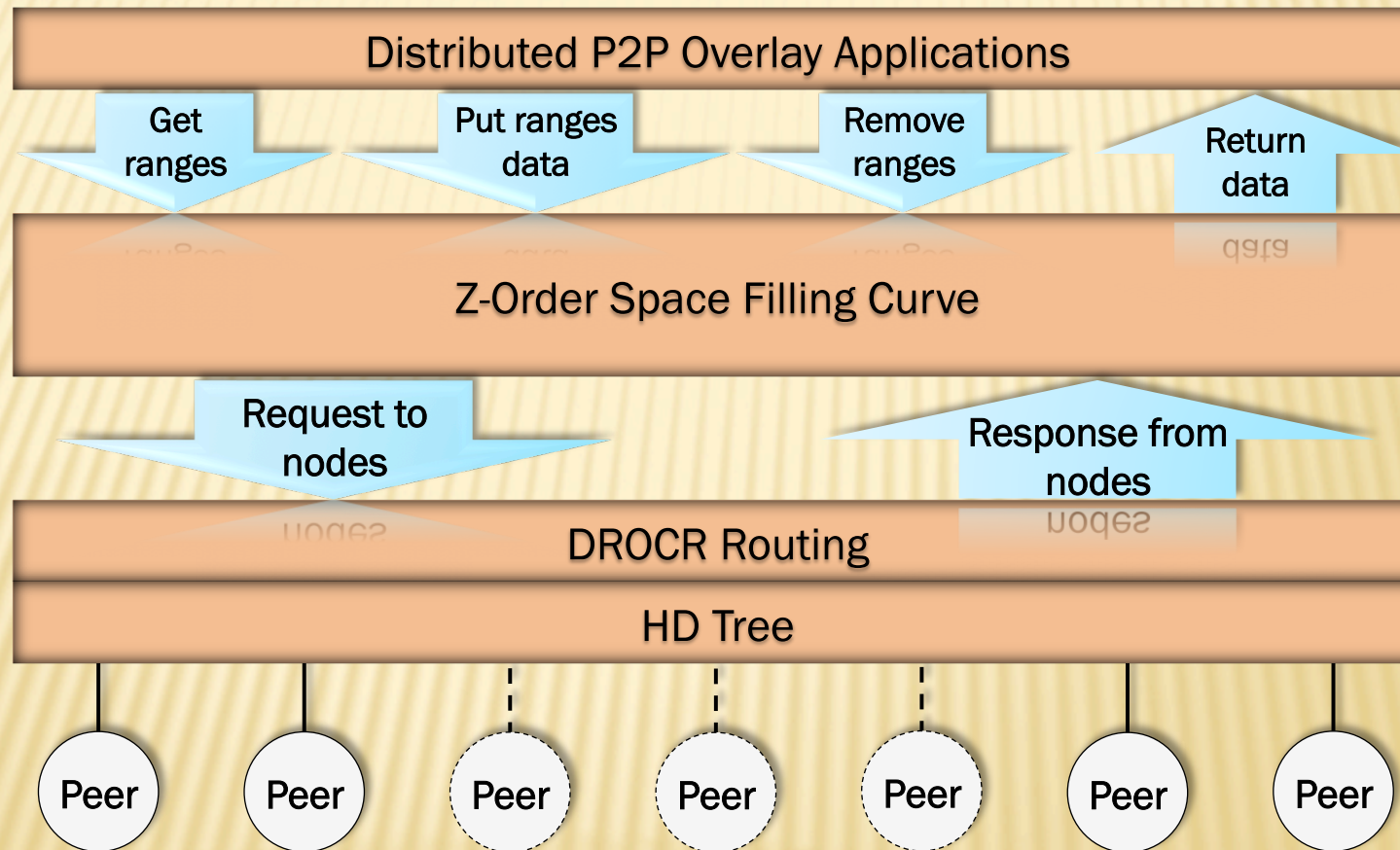
HR: 2HP/2DP/2HC/2DC routing

Tree: routing in the equivalent tree structure

ERROR RESILIENT ROUTING

- ✗ Match exists between SRC and DST codes
 - + Try all possible options that can extend existing match towards DST code
 - + Try all options towards DST node
 - + Try SDP & SDC
- ✗ No match can be found between SRC and DST codes
 - + Try all possible options towards DST node
 - + Try SDP & SDC

SUPPORTING MDR QUERIES IN HD TREE



SUPPORTING MDR QUERIES IN HD TREE

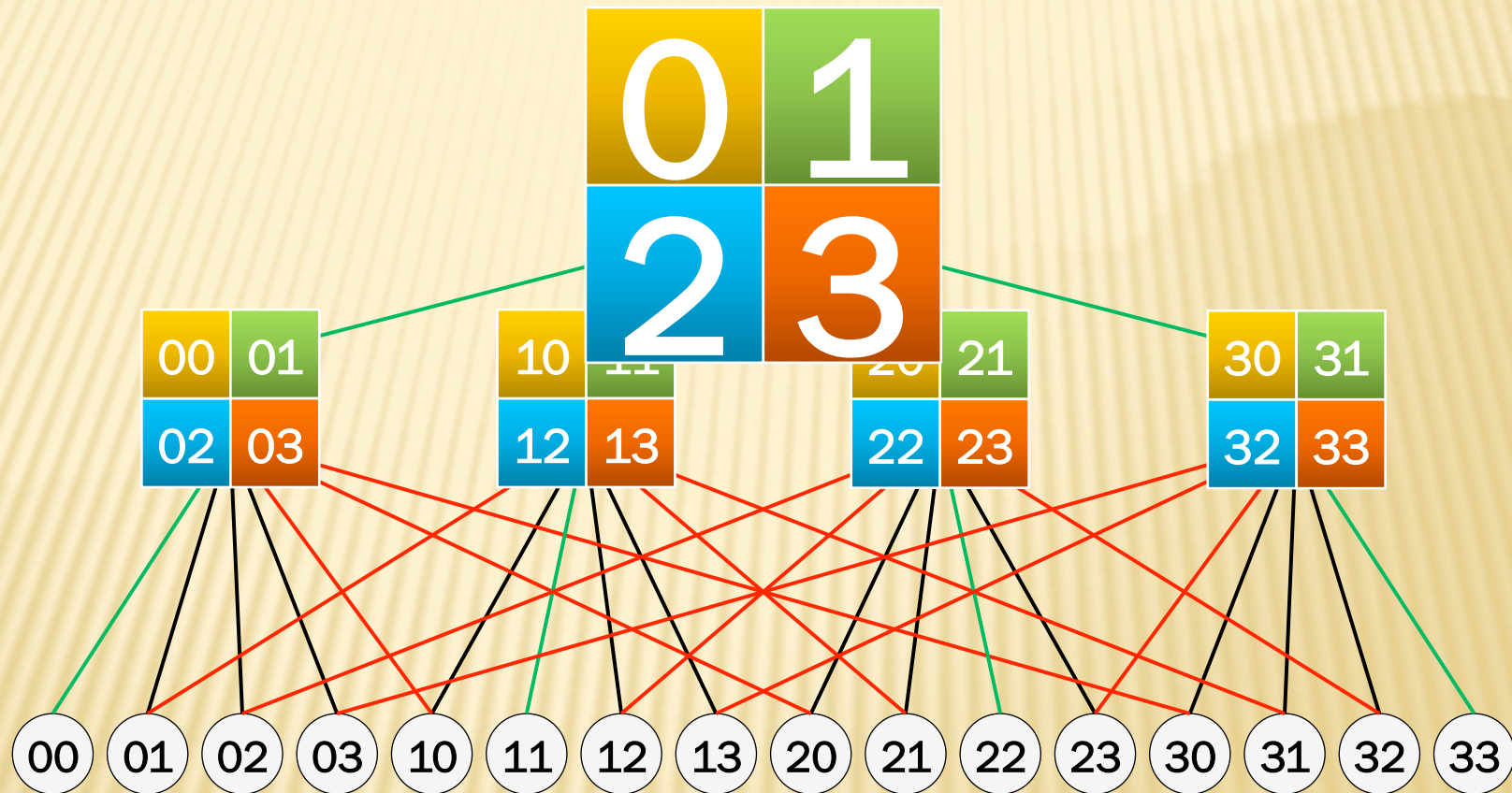
0	1
2	3

1st order of recursive decomposition

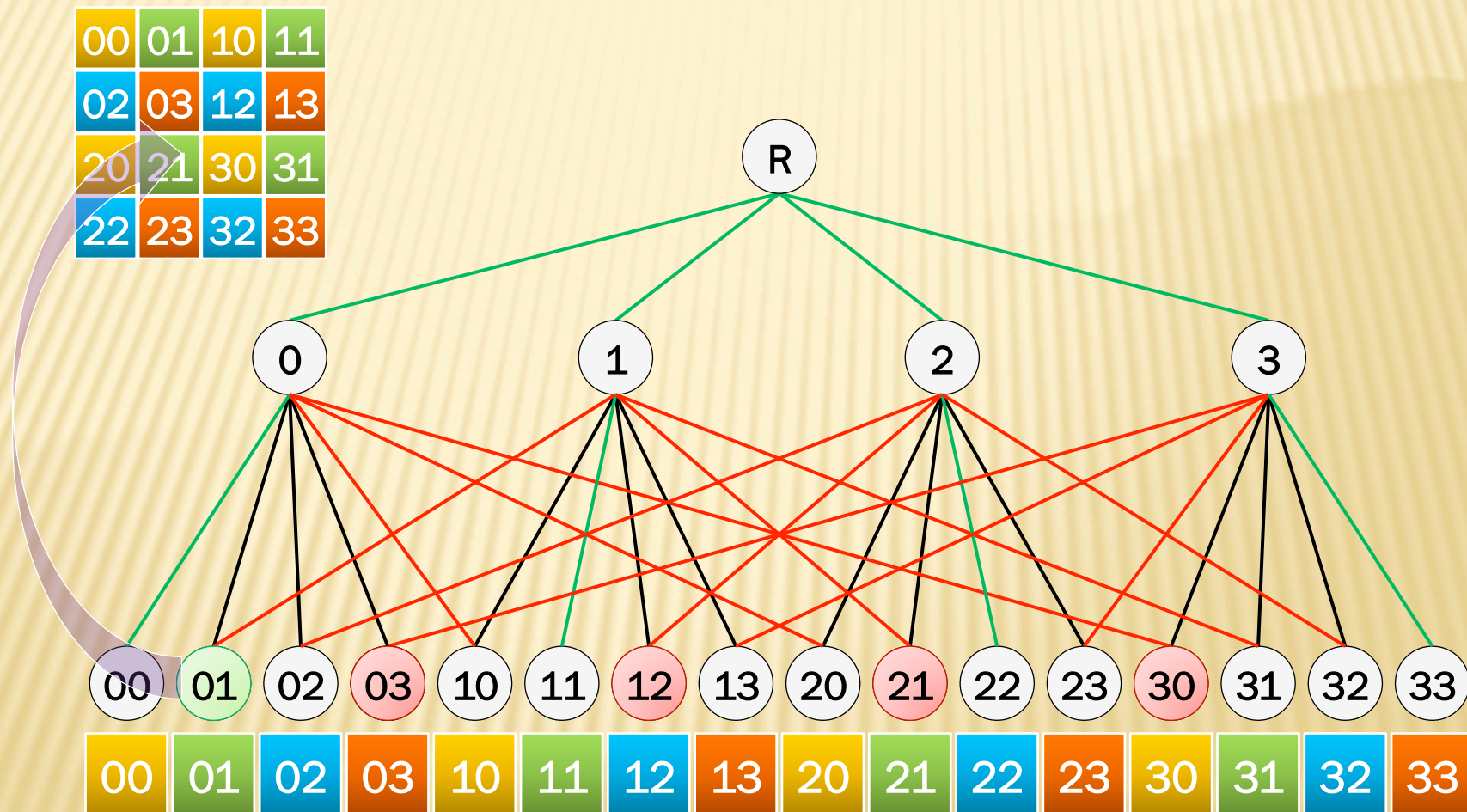
00	01	10	11
02	03	12	13
20	21	30	31
22	23	32	33

2nd order of recursive decomposition

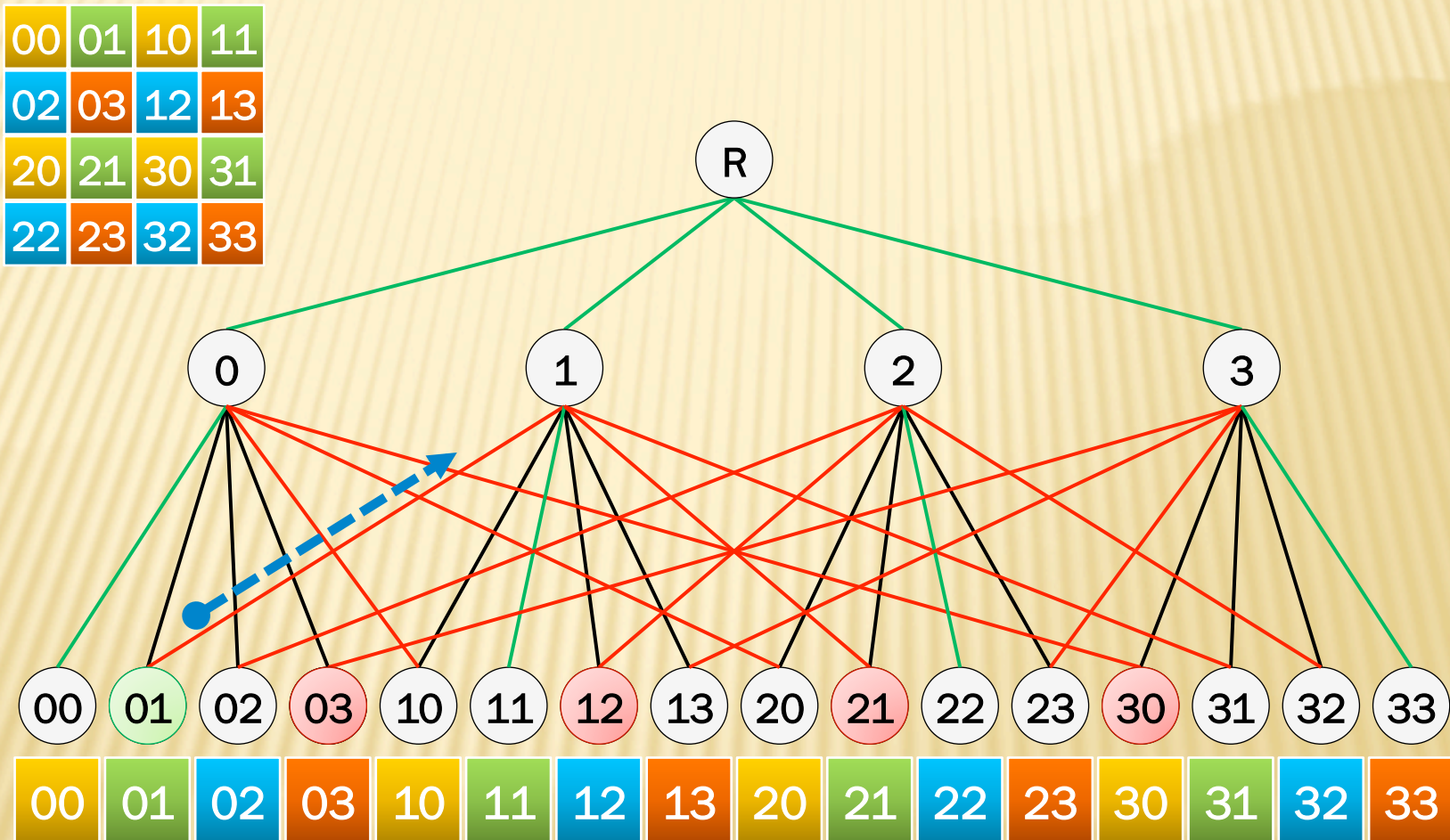
SUPPORTING MDR QUERIES IN HD TREE



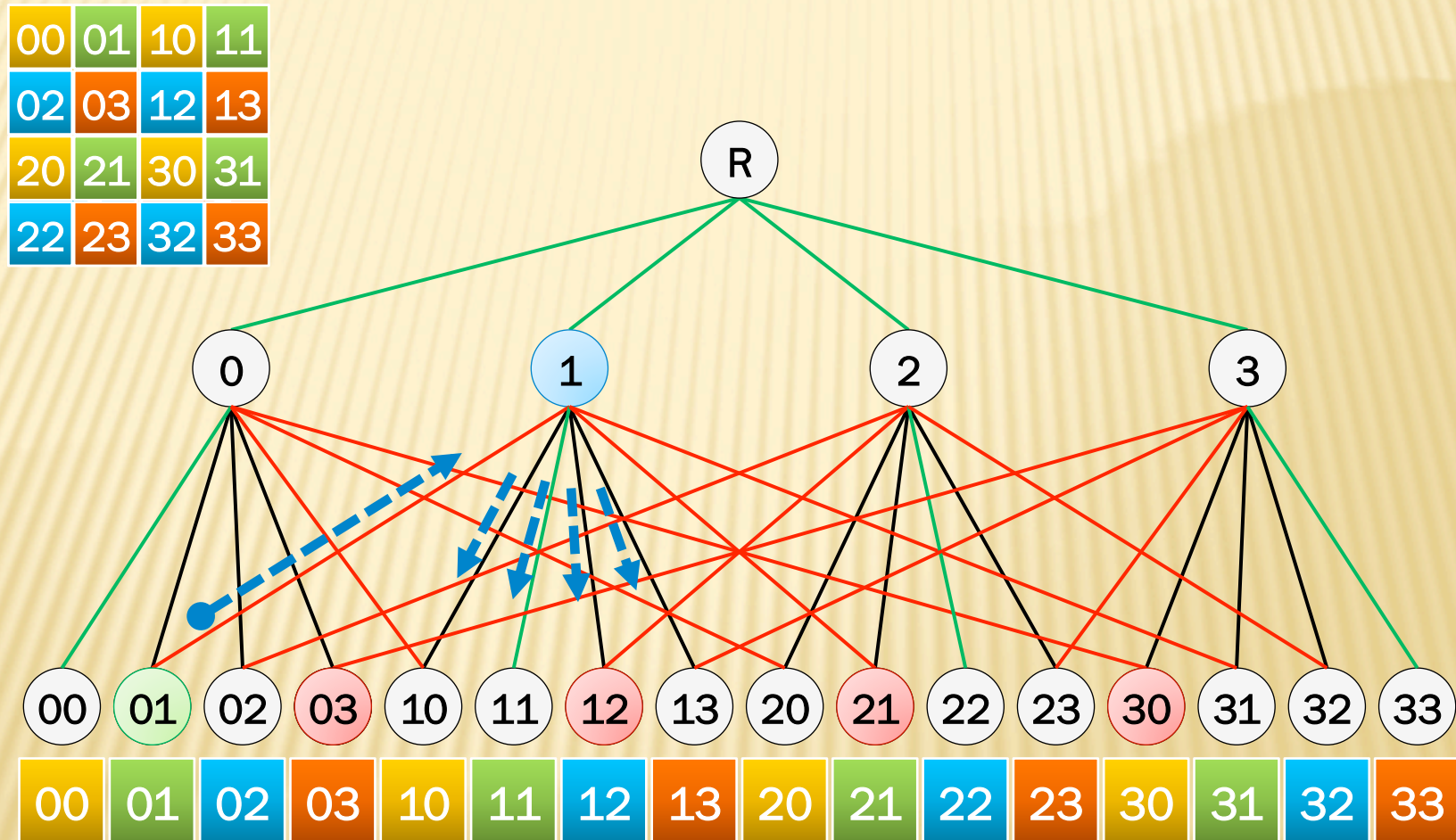
MDR QUERIES – D2H QUERY



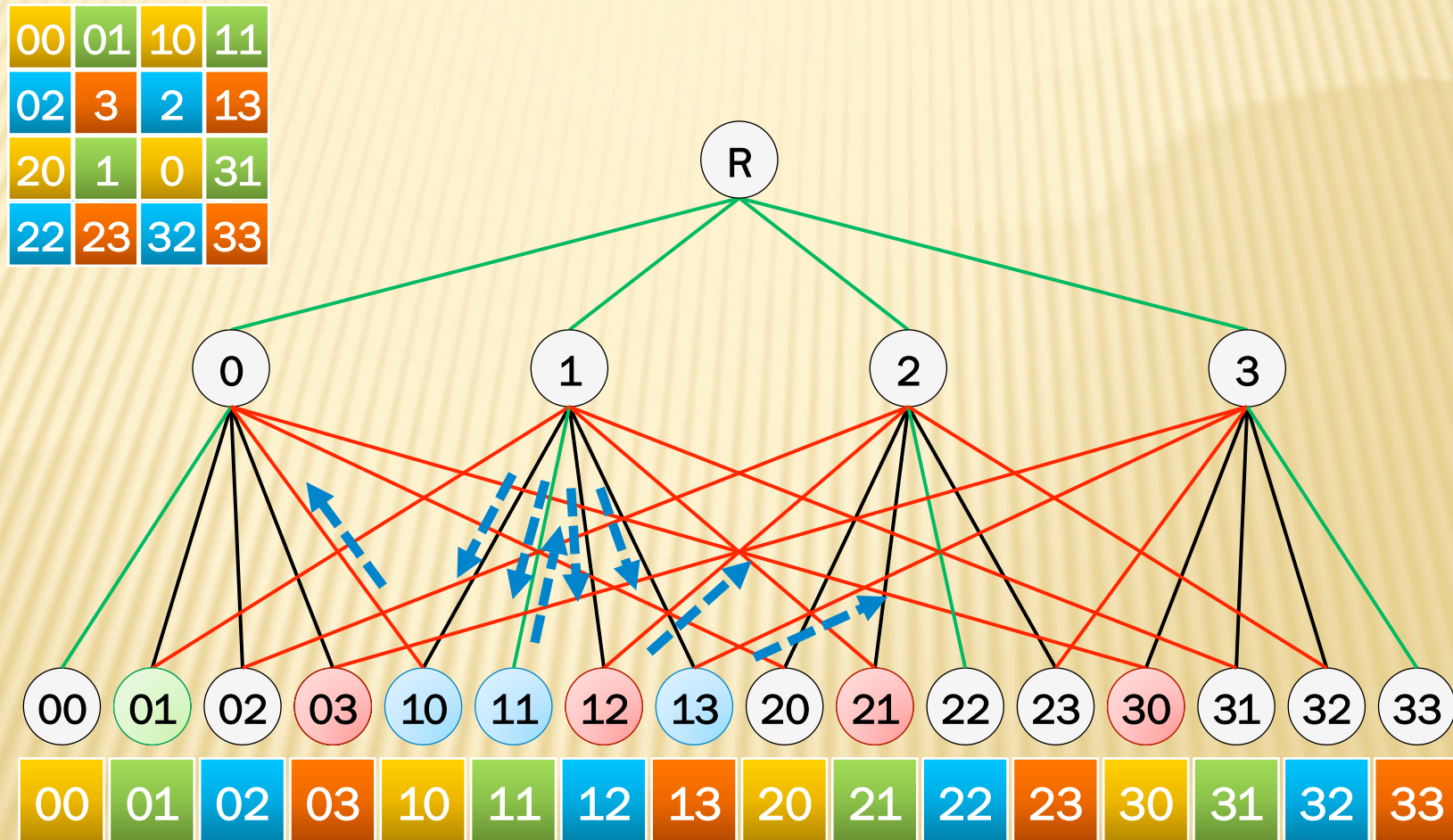
מרכז המידע והסיוע - 052-5252525



MDR QUERIES – D2H QUERY

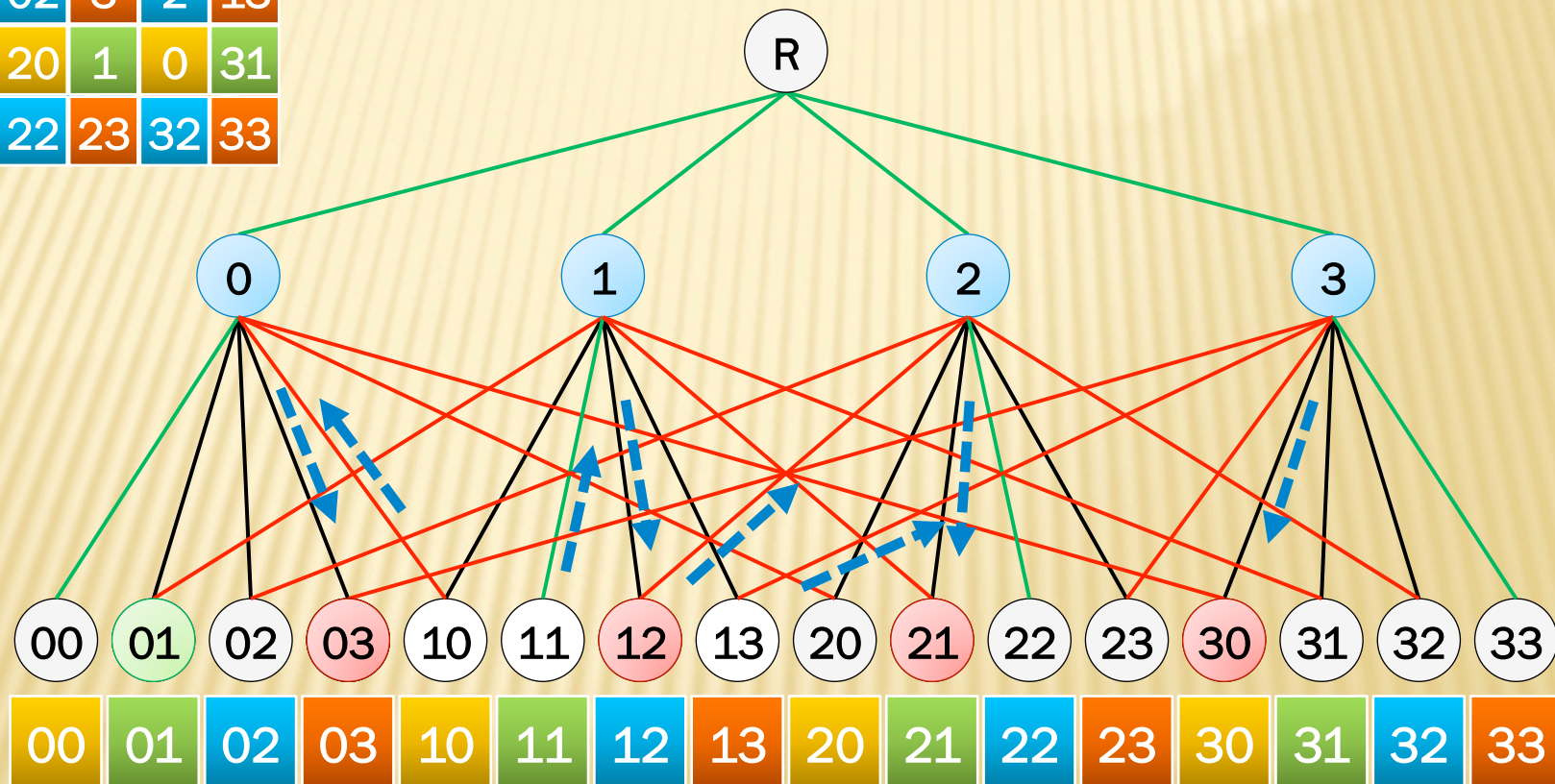


MDR QUERIES – D2H QUERY



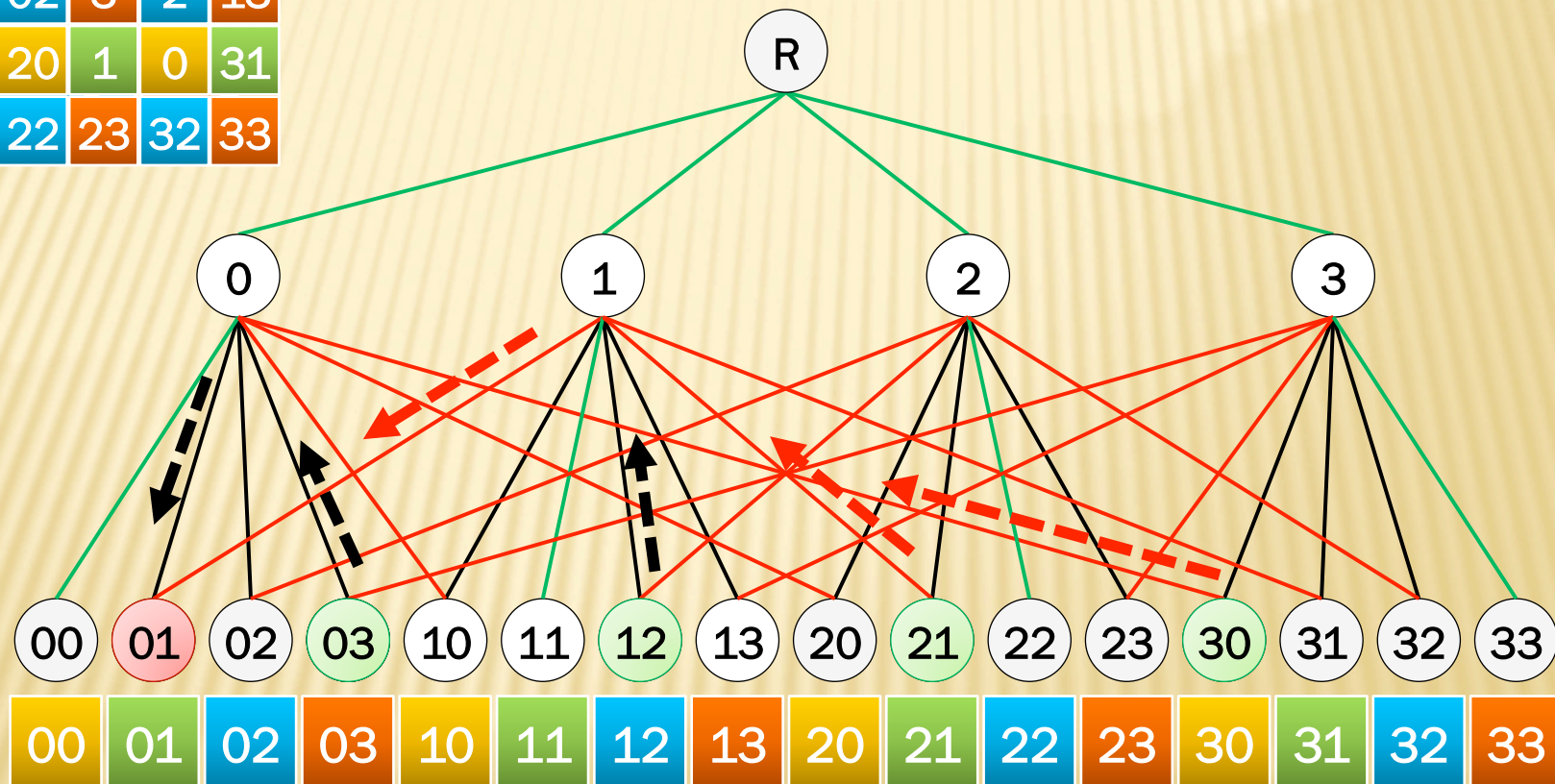
MDR QUERIES – D2H QUERY

00	01	10	11
02	3	2	13
20	1	0	31
22	23	32	33

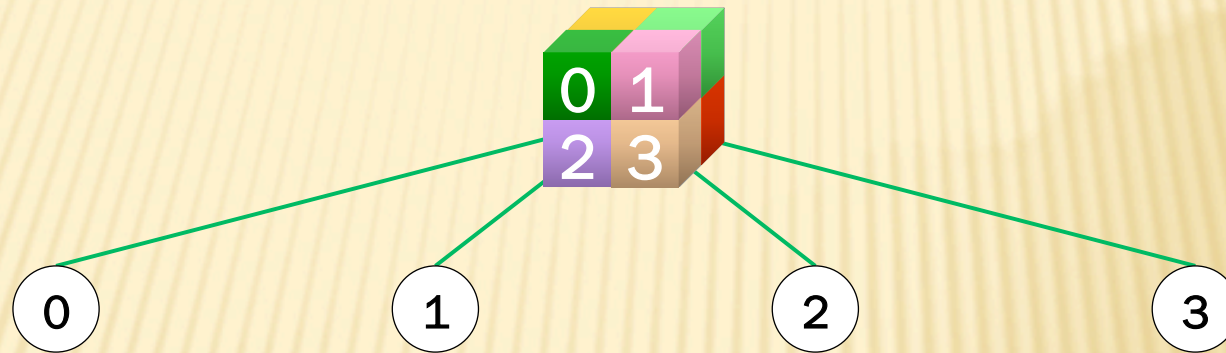


MDR QUERIES – DROCR RESPONSE

00	01	10	11
02	3	2	13
20	1	0	31
22	23	32	33

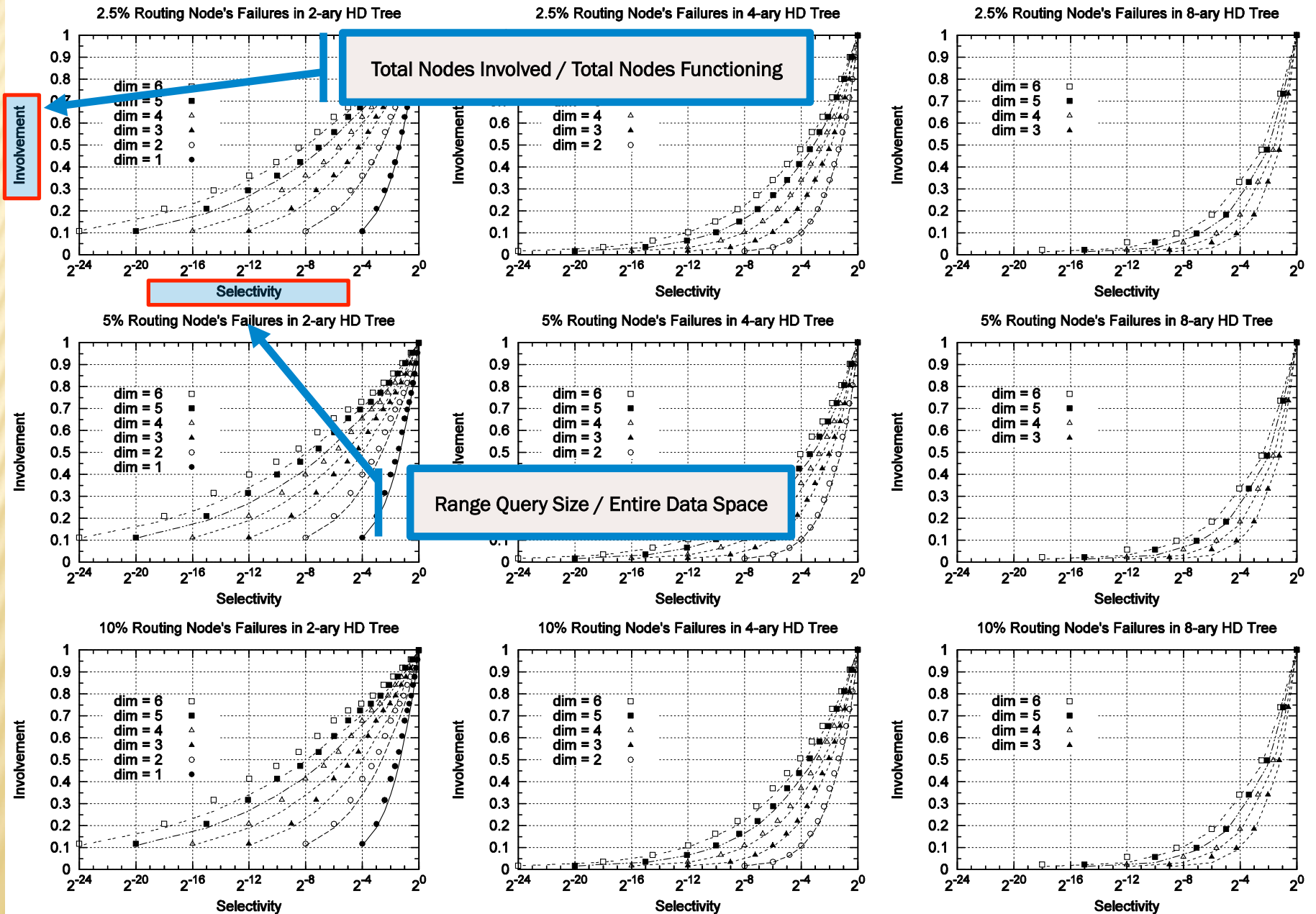


SUPPORTING MDR QUERIES IN HD TREE



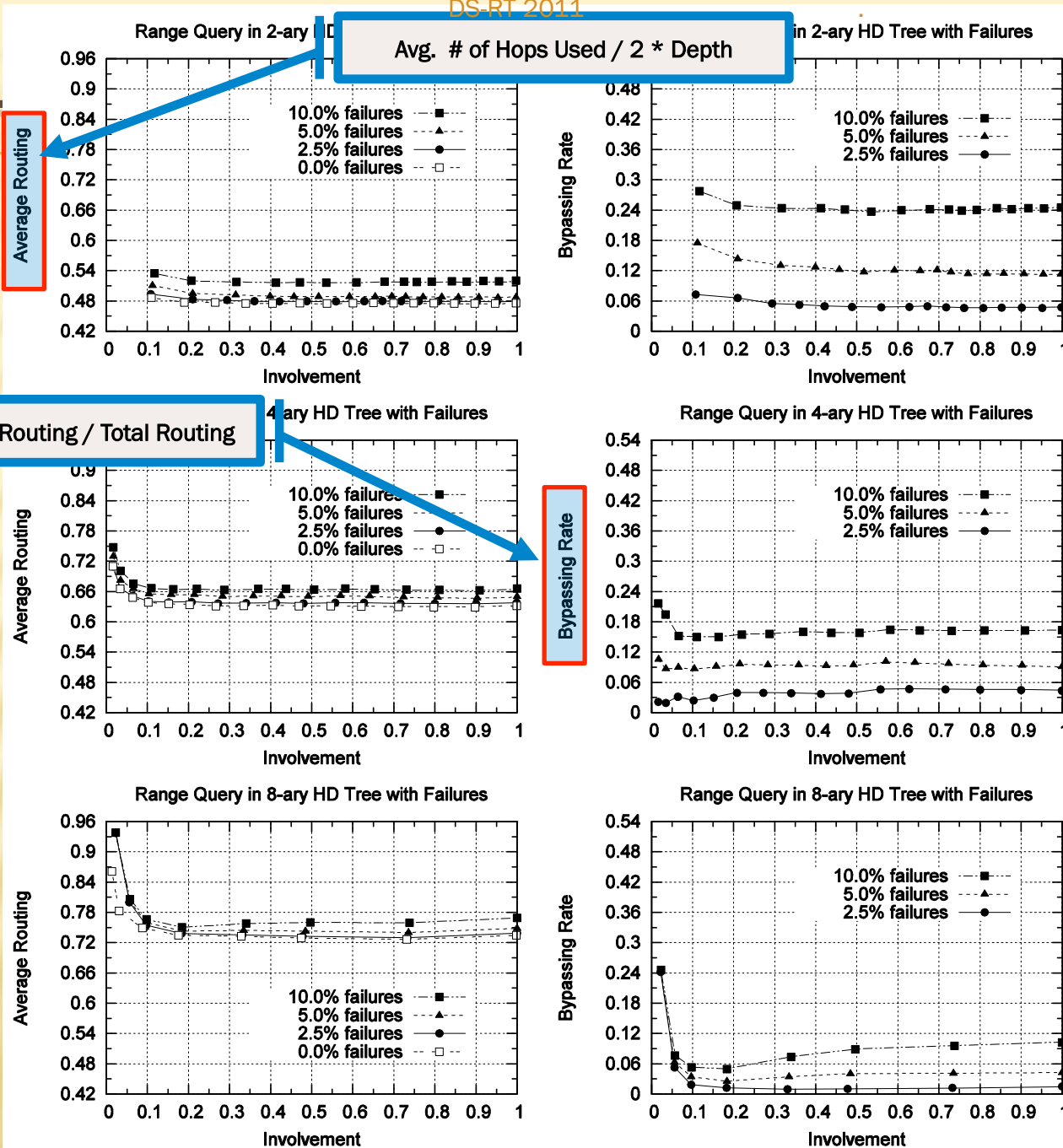
SIMULATION RESULTS

- ✗ Simulator - ns2
- ✗ Network Size
 - + 2-ary HD tree of height 9 ($n = 1023$)
 - + 4-ary HD tree of height 5 ($n = 1365$)
 - + 8-ary HD tree of height 3 ($n = 585$)
- ✗ Multi-dimensional Range Queries
 - + Z-Order SFC & DROCR
 - + AVG # of nodes affected / total # of nodes in the system
 - + Dimensionality(dim): 1 – 6
 - + Selectivity
 - ✗ from $1/16^{\text{dim}}$ to $16^{\text{dim}}/16^{\text{dim}}$ in 2 & 4-ary HD tree
 - ✗ from $1/8^{\text{dim}}$ to $8^{\text{dim}}/8^{\text{dim}}$ in 8-ary HD tree



SIMUL

DS-RT 2011



CONCLUSIONS

- ✗ HD Tree is the first data structure that is able to give a complete view of the system state when processing MDRQ at different levels of selectivity.
- ✗ In a k-ary HD Tree
 - + Root can survive $(k-1)$ point failures
 - + Internal nodes can survive $(2k-1)$ point failures
 - + Leaf nodes can survive 1 point failure
- ✗ Routing in HD Tree outperforms routing in the equivalent tree structure
- ✗ MDRQ can be optimized in the HD Tree network because it is
 - + direct mapping in nature
 - + greedy and optimal routing strategies
- ✗ 10% of routing nodes' failures do not cause significant performance variations in HD Tree.
- ✗ Further improvement is possible if all possible routing options can be explored in the by-passing routing mode.

REFERENCES

- [1] J.L. Bentley, "Multidimensional Binary Search Trees Used for Associative Searching" Commun. of ACM, Vol18, No. 9, (Sep 1975), pp 509-517
- [2] R.A. Finkel, and J.L. Bentley, "Quad trees: a data structure for retrieval on composite key." Acta Informatica 4, 1 (1974), pp 1-9.
- [3] Hanan, "The Quadtree and Related Hierarchical Data Structures ", ACM Computing Surveys (CSUR), Volume 16 , Issue 2, (June 1984) pp 187-260
- [4] G.M. Morton, "A Computer Oriented Geodetic Data Base and a New Technique in File Sequencing," technical report, Ottawa, Canada, IBM Ltd. 1966.
- [5] H. Jagadish. "Linear clustering of objects with multiple Attributes." In Proc. SIGMOD, Vol 17, Issue 2, 1990, pp 332-342
- [6] J. Orenstein and T. Merrett. "A class of data structures for associative searching." Proc. PODS, 1984, pp 181-190
- [7] James Aspnes, Gauri Shah "Skip graphs" Fourteenth Annual ACM Symposium on Discrete Algorithms (SODA), January 2003, pp. 384-393..
- [8] S. Ratnasamy et al., "A scalable content-addressable network." Proc. ACM SIGCOMM, 2001, Vol. 31, No. 4. (Oct. 2001) pp. 161-172.
- [9] I. Stoica, R. Morris et al., "Chord: a scalable peer-to-peer lookup protocol for Internet applications". IEEE/ACM Trans. Vol. 11, No. (Feb, 2003), pp. 17-32.
- [10] P. Ganesan, B. Yang, H. Garcia-Molina, "One torus to rule them all: multi-dimensional queries in P2P systems" In WebDB '04: Proceedings of the 7th International Workshop on the Web and Databases (2004) . pp: 19 - 24
- [12] Yanfeng Shu, Beng Chin Ooi, Kian-lee Tan "Supporting Multi-dimensional Range Queries in Peer-to-Peer Systems" In Fifth IEEE International Conference on Peer-to-Peer Computing, 2005. P2P 2005. pp 173 - 180
- [13] Chi Zhang, Arvind Krishnamurthy, Randolph Y Wang "SkipIndex: Towards a Scalable Peer-to-Peer Index Service for High Dimensional Data" Vol. TR-703-04 (May 2004). Princeton University Computer Science Department
- [14] Cristina Schmidt, Manish Parashar "Flexible information discovery in decentralized distributed systems" Proceedings of the 12th IEEE International Symposium on High Performance Distributed Computing (HPDC'03) pp 226 - 235
- [15] N. Harvey, M.B. Jones, S. Saroiu, M. Theimer, A. Wolman, "SkipNet: a scalable overlay network with practical locality properties" in: Proceedings of the 4th USENIX Symposium on Internet Technologies and Systems USITS'03, March 2003.
- [16] K. Aberer, P. Cudre-Mauroux, A. Datta, Z. Despotovic, M. Hauswirth, M. Ponceva, R. Schmidt, "P-Grid: a self-organizing structured P2P system" ACM SIGMOD Record 32 (3) (2003) 29-33.
- [17] A. Bharambe, M. Agrawal, S. Seshan, "Mercury: supporting scalable multi-attribute range queries" SIGCOMM'04, August 30-September 3, 2004.
- [18] J. Risson, T. Moors, "Survey of research towards robust peer-to-peer networks: search methods" Computer Networks, Vol50 No17, (Dec 2006), pp 3485-3521