Communication Environment of Next Generation Supercomputers and the Theory of Spatially Embedded Complex Networks

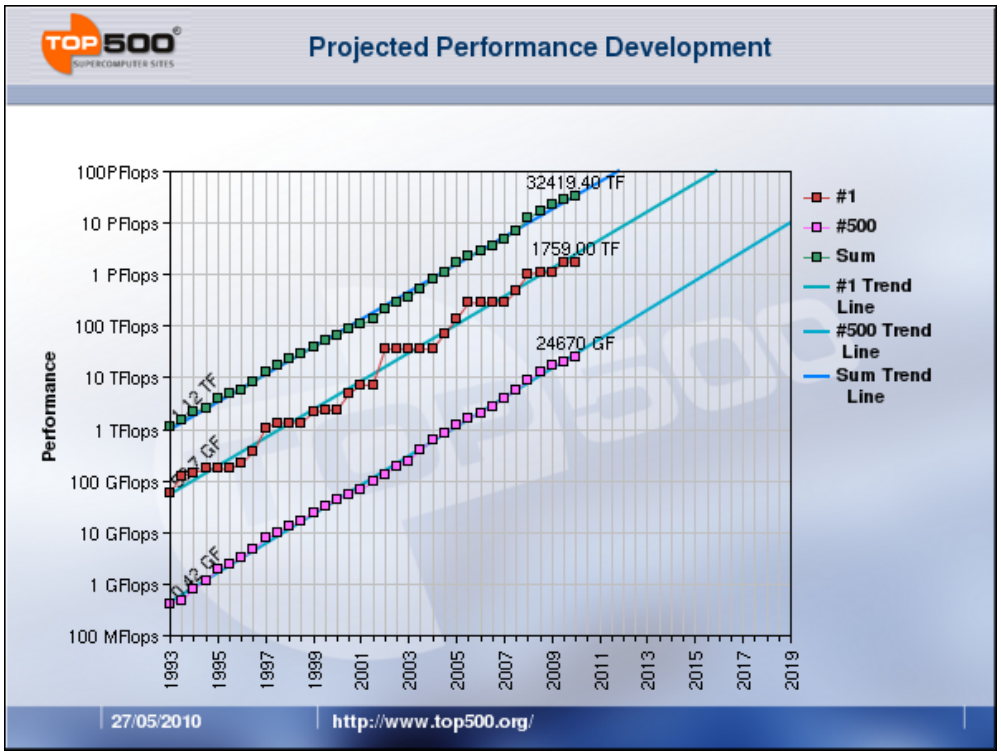
A.P.Demichev, V.I.Ilyin, A.P.Kryukov

NRC "Kurchatov Institute" and SINP MSU

The 5th International Conference "Distributed Computing and Grid-technologies in Science and Education"
Grid'2012

Exascale Architecture Challenges

- After the Petaflop (10^{15} FLOPS) performance barrier have broken, the HPC community is exploring development efforts for breaking the Exaflop barrier (10^{18} FLOPS).
  - the 1st Exaflop system is estimated to be built between 2018 and 2020
  - a number of technical challenges to constructing an Exascale computing system



Exascale Architecture Challenges: Interconnection networks

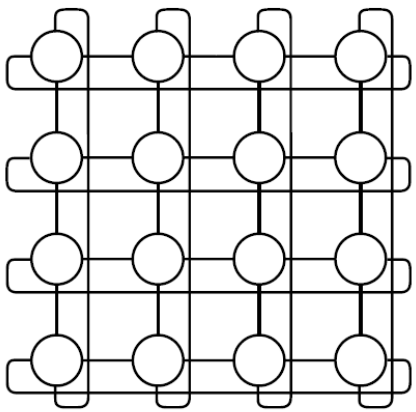
- Interconnection network is one of the key hardware component that define the level of the architecture scalability
  - e.g., IAA Interconnection Networks Workshop & refs therein
- Among others, Exascale systems are likely to require the following from interconnection networks:
  - Scalability: greater than 100,000 endpoints
  - High Message Throughput: greater than 100 million messages/s for MPI
  - Low Latency: Maintain approximately 1 μs latency across system
  - High Reliability: less than 10^{-23} unrecovered bit error rates

Some Basics on Interconnection Networks

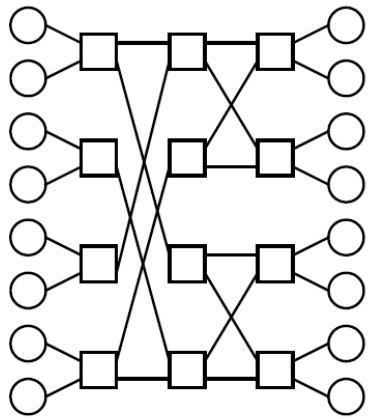
- The three core aspects of interconnection networks:
  - Topology
  - Flow control
  - Routing
- In this communication we consider mainly the topology of interconnection networks
- Two other aspects are very important but out of the scope of the current presentation
  - Selecting a suitable topology is vital to the design of a network as the routing and flow control mechanisms will rely heavily on its properties and the physical implementation must be within the means of the fabrication process in terms of cost and technology

Types of Network Topologies

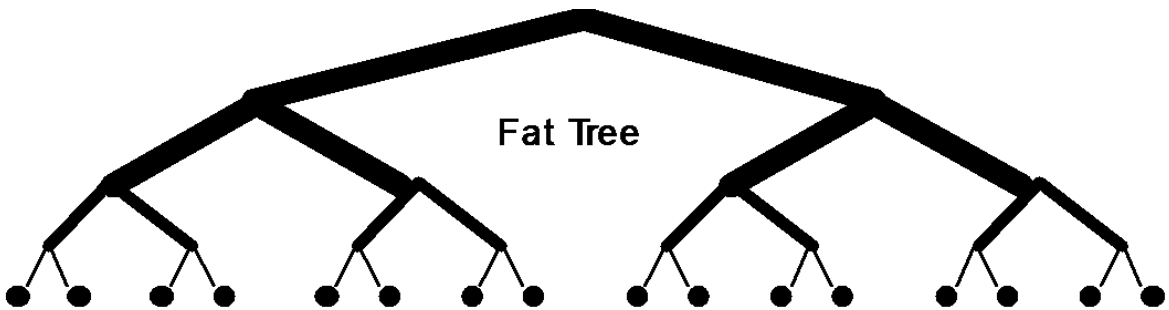
- a direct network: $\forall$ node = terminal = source + sink for messages + switch to route incoming messages (ex. (a)); an indirect network contains non-terminal nodes that are used just for switching (ex. (b)); one more example of topology - "fat tree network"



(a) Torus (4-ary 2-cube)

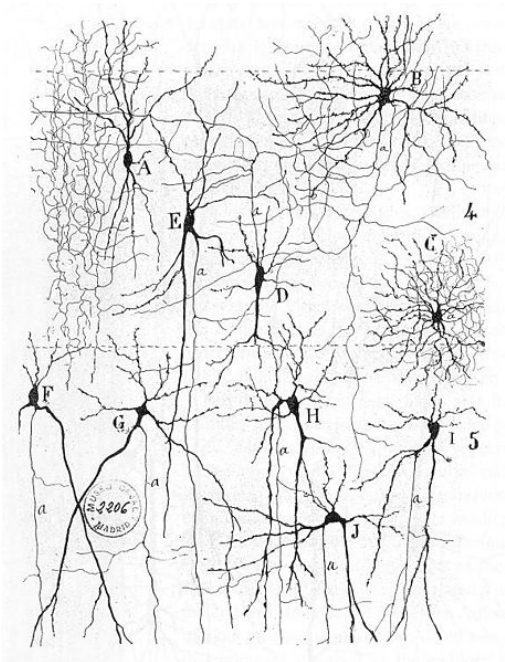


(b) Butterfly (2-ary 3-fly)



Advanced Approach to the Interconnection for the Next Generation Supercomputers: Neurocomputers

- Neurocomputers $\Rightarrow$ a new class of parallel, distributed-information processors that mimic the functioning of the human brain, including its capabilities for self-organization and learning
- The theory of spatially embedded complex networks is a powerful tool for the development of principles of communication environment design for neurocomputers and hybrid (bioelectronic) systems
  - The concepts are still being researched, but in the near future, it is expected that neurocomputers, even though they are still considerably simpler in design than animal brains, should be capable of simple pattern recognition tasks such as handwriting recognition.



A Choice of the Topology for further Investigations

- Indirect networks have their own merits but they are hardly scalable to the huge number of nodes
  - We consider only direct networks
- Usually the interconnection networks have regular structure; advatages:
  - in general, highly reliable with respect to faults of some nodes/channels ($\exists$ alternative routes),
  - hypercube with dimensionality enough for a given number of nodes provides short mean path length between the nodes (mean distance)
  - for 3D-simulations the structure of the tasks optimally maps to the natural 3D lattice of the computational nodes

Limitations of Regular Interconnection Network Topologies for Exascale Computing ($> 10^5$ compute nodes)

- low dimensional lattices $\Longrightarrow$ high mean distance
- high dimensional lattices (k-ary n-meshes, hypecubes) (D compareble with number of nodes) $\Longrightarrow$ difficult for implementation, huge length of physical channels
- fat tree networks $\Longrightarrow$ huge number of switches + necessity of huge bandwidth in the origin of tree root

Project goals

- Developing an alternative approaches to the design of the communication environment architecture for exascale supercomputers, namely
  - an approach to design of interconnection networks of more general type: with inhomogeneous structure
- Currently the project is in its initial stage:
  - formulation of basic principles
  - choice of methods for designing and investigation of irregular networks
  - development of simulation toolkit
- No ready prescriptions yet

Strategy of the Approach

- Preferably the network should preserve the structure of a 3D regular lattice
  - e.g., for numerical simulation of 3D-objects
- This basic structure should have additional links so that general type computing tasks could be effectively processed as well
  - $\Rightarrow$ small mean distance, optimal navigation and routing in the network
- Features of these new links should depend on variable parameters
  - under variation of these parameters the network should vary from regular to fully random

Importance of Small Average Path Length Between Nodes: A Simple Model for Message Flow

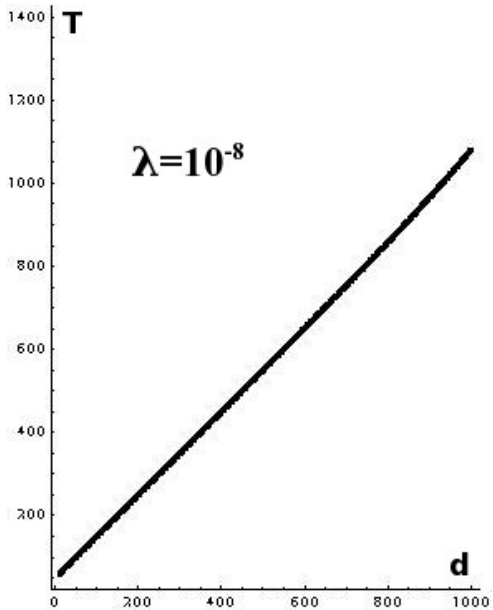
- Each D-torus dimension is modelled as a queueing system of the $M/G/1$ type
  - because of possible blocking while trasition of multiple messages in the network, service time is supposed to be arbitrary
- Using two-moment approximation in the theory of queueing systems + a number of simplifying approximations
  - e.g., random traffic pattern
- we managed to obtain a closed equations for time lattency, in particular for 1D-torus:

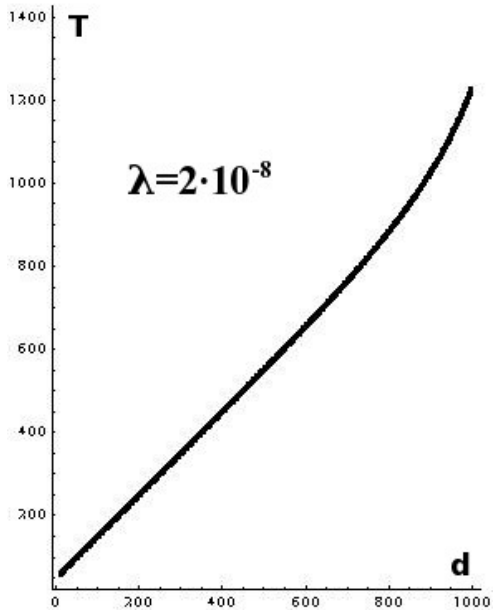
$$\bar{T}_1 = M + \bar{k}\left\{1 + \frac{\lambda_g \bar{k}^2 \bar{T}_1^3 [1 + (\bar{T}_1 - M)^2/\bar{T}_1^2]}{8(1 - \lambda_g \bar{k}\,\bar{T}_1/2)}\right\} .$$

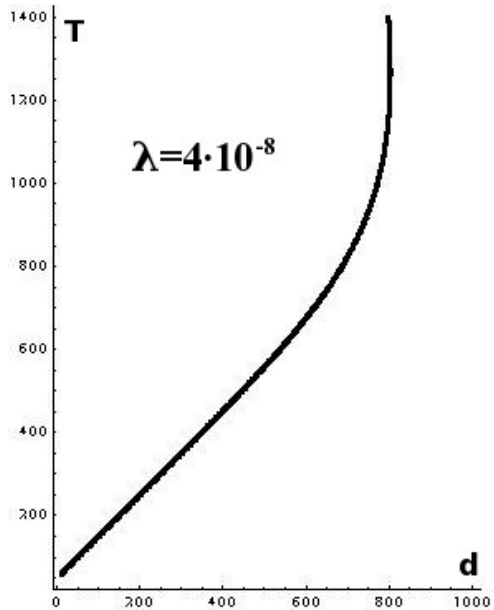
- due to simplicity and essential approximation the model works correctly only for small enough message generation rate λ_g

Importance of Small Average Path Length Between Nodes: A Simple Model for Message Flow (2)

- The model predicts a sharp increase of the latency with increasing the average path length even for relatively low values of λ_g



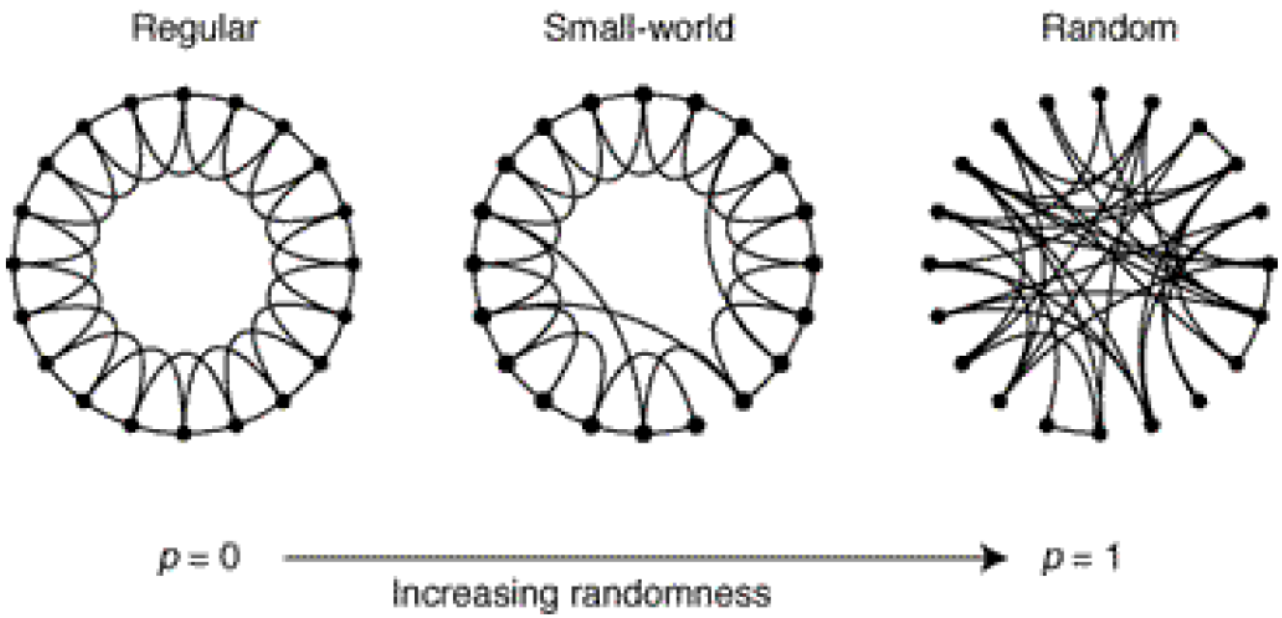




- Thus it is important to develop an approach to disigning networks with small average path length but preserving the lattice structure

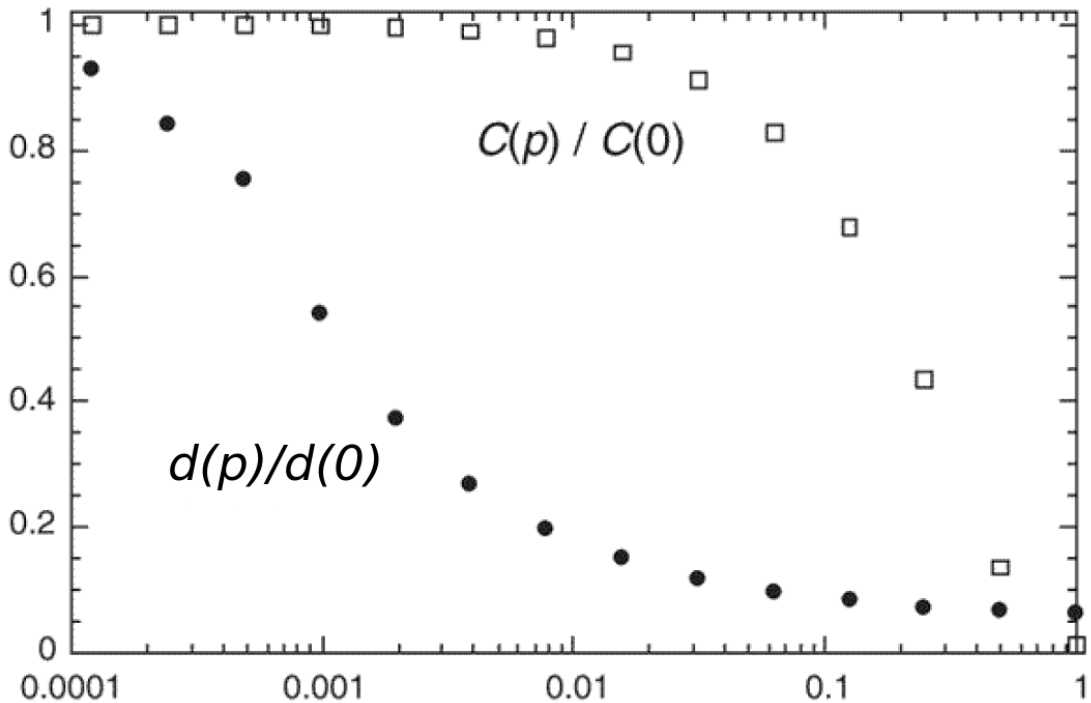
Basic idea: Using the Small-world Phenomenon in Complex Networks

- Algorithm of construction:
  - Define a number of nodes N and a probability $0 \leq p \leq 1$ ($N \gg K \gg \ln N \gg 1$)
  - Construct a lattice with K neighbors
  - $\forall$ node take the edge (n_i, n_j) $(i < j)$ and with the probability p rewire to the node n_k (chosen with uniform probability); selfloops and double edges are excluded



Small-World Model: Basic Properties

- Of principal importance - combination:
  - high clasterization — similar to regular lattices
  - small average length — similar to random graphs
- As a result of the algorithm application there appear $\sim pNK/2$ irregular edges - shorcuts
- Varying p from 0 to 1 produces interpolation between regular lattice and fully irregular random graphs

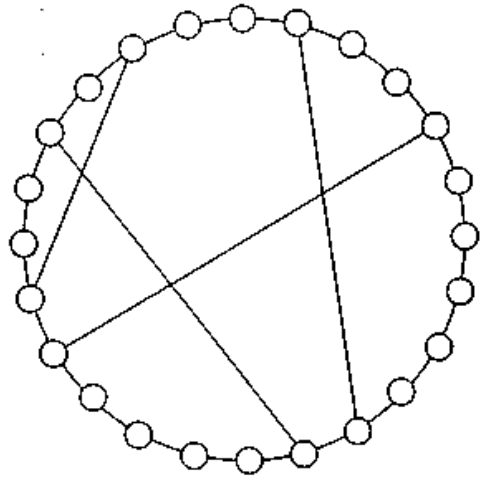


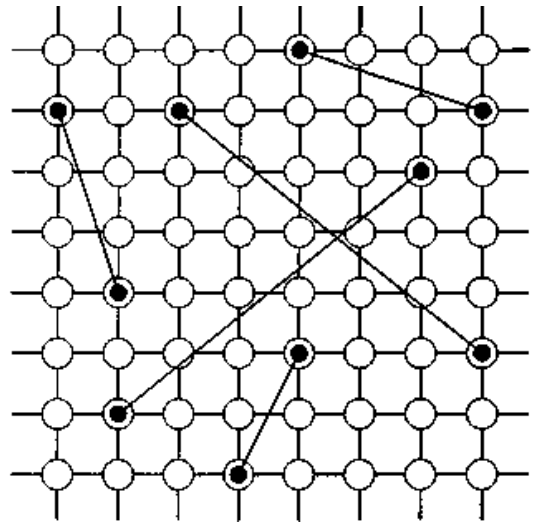
Topology for Exascale: Spatially Embedded Small-World Network Preserving Lattice Structure

- Staring point is D-dimensional regular lattice with $N = L^D$ nodes
- $\forall$ node a shortcut is added with probability p (pN additional shortcuts on average)
- Since the shortcuts have to be physically realized there is a cost associated with their length; a way to model this is to add links with a probability

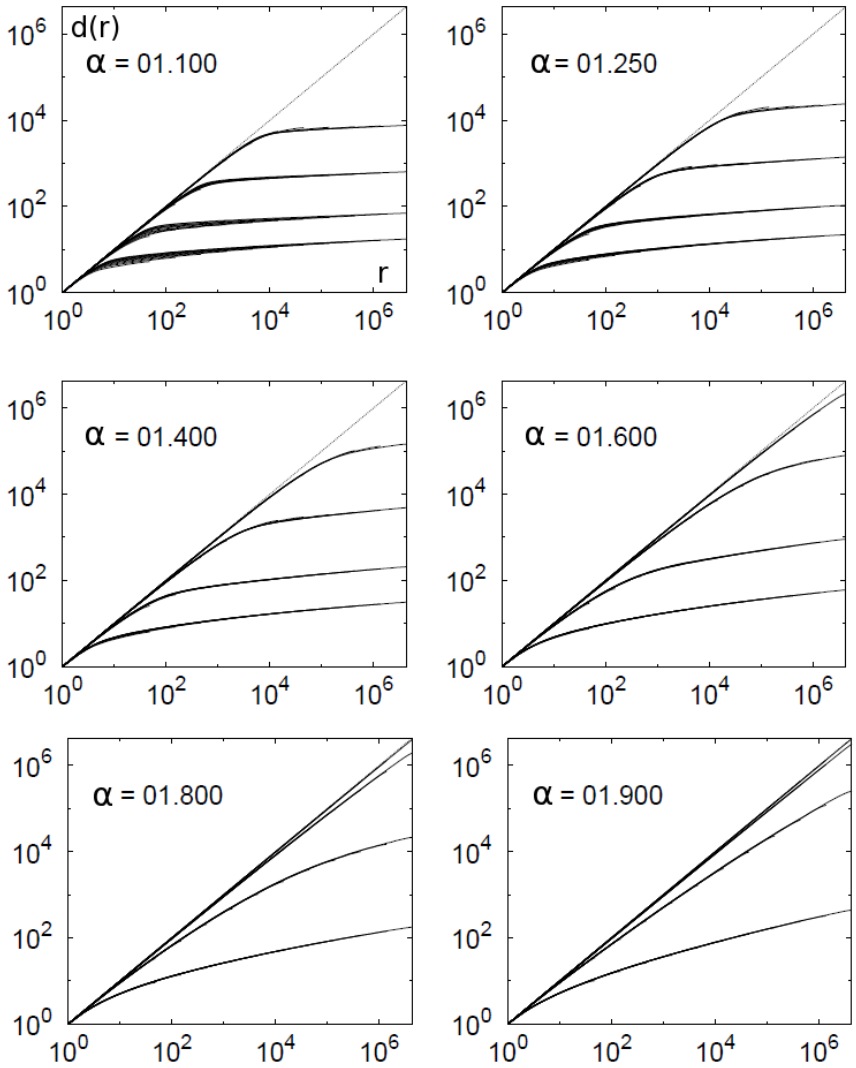
$$P(r) \sim r^{-\alpha}$$

- Such networks considered in the literature but not in the context of supercomputer interconnection networks



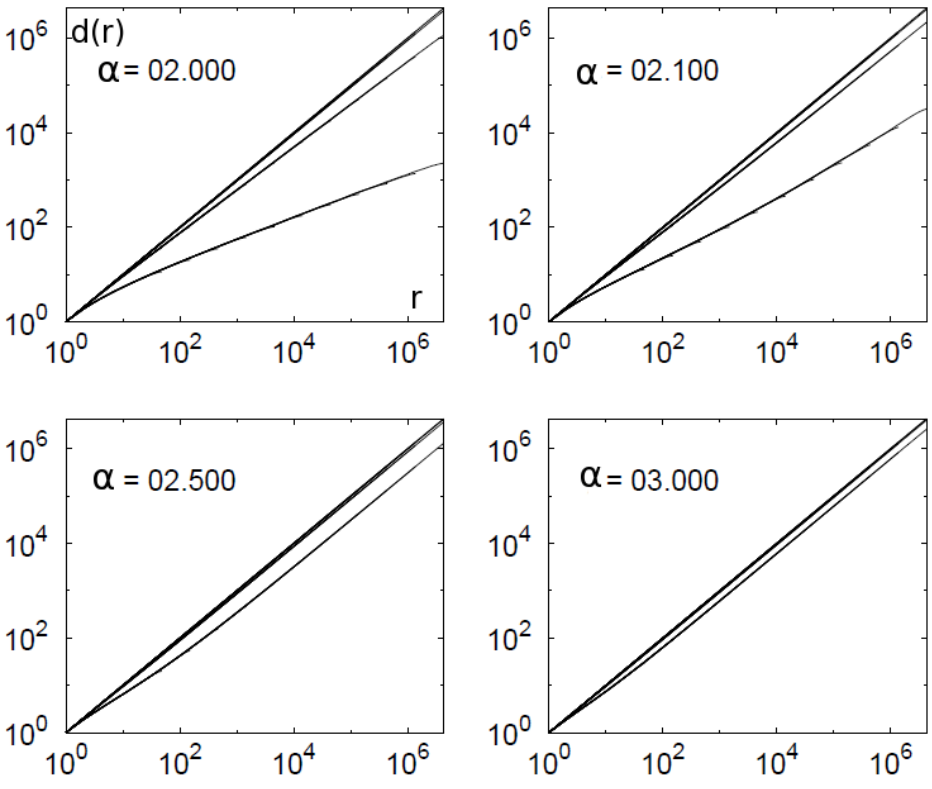


Numerical simulations for $D = 1$ and $\alpha < 2$



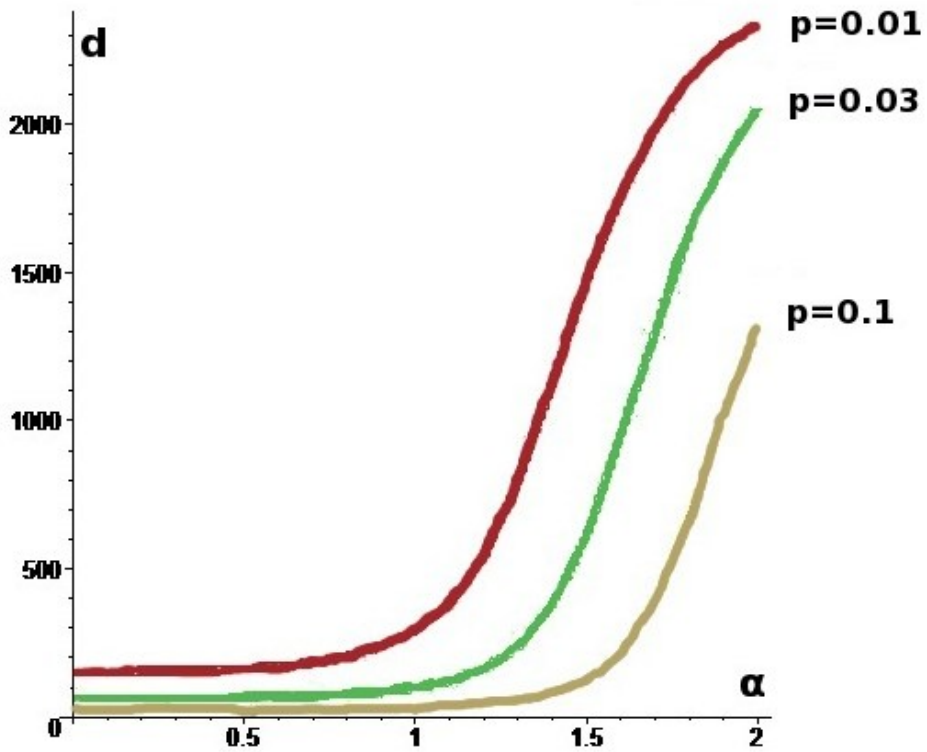
- $d(r)$ для $1 < \alpha < 2$
- Shortcut densities: $p = 10^{-3},\ 10^{-2},\ 10^{-1},\ 1$
- $\forall\ \alpha < 2$, d at first grows linearly with r, then at some characteristic $r = \xi$ (depending on p and α) grows much slower: $r^{\theta_s},\ \theta_s < 1$.

Numerical simulations for $D = 1$ and $\alpha > 2$



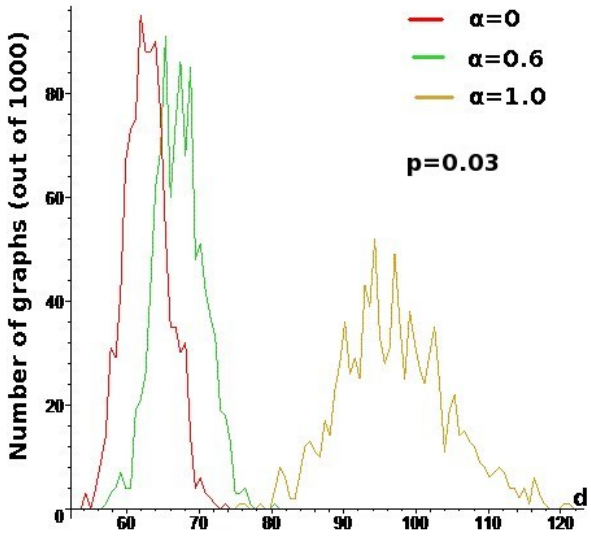
- No Small-World effect

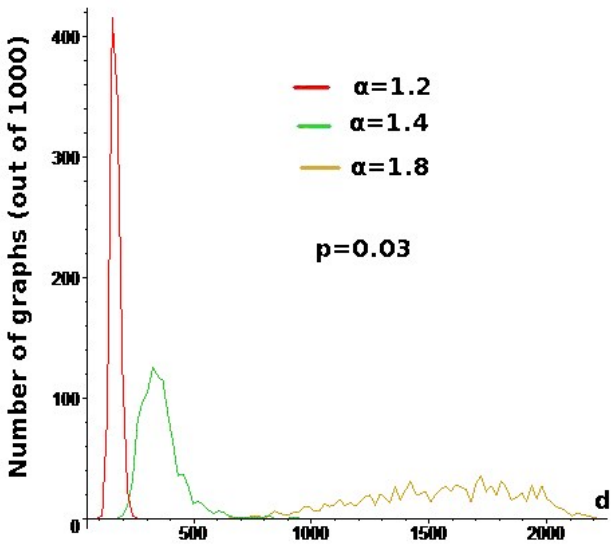
Dependence of the average path length on the shortcut length distribution $P(r) \sim r^{-\alpha}$ and density p



- Average path length grows with growing α
  - numerical simulation is thanks to S.Polyakov (SINP MSU)

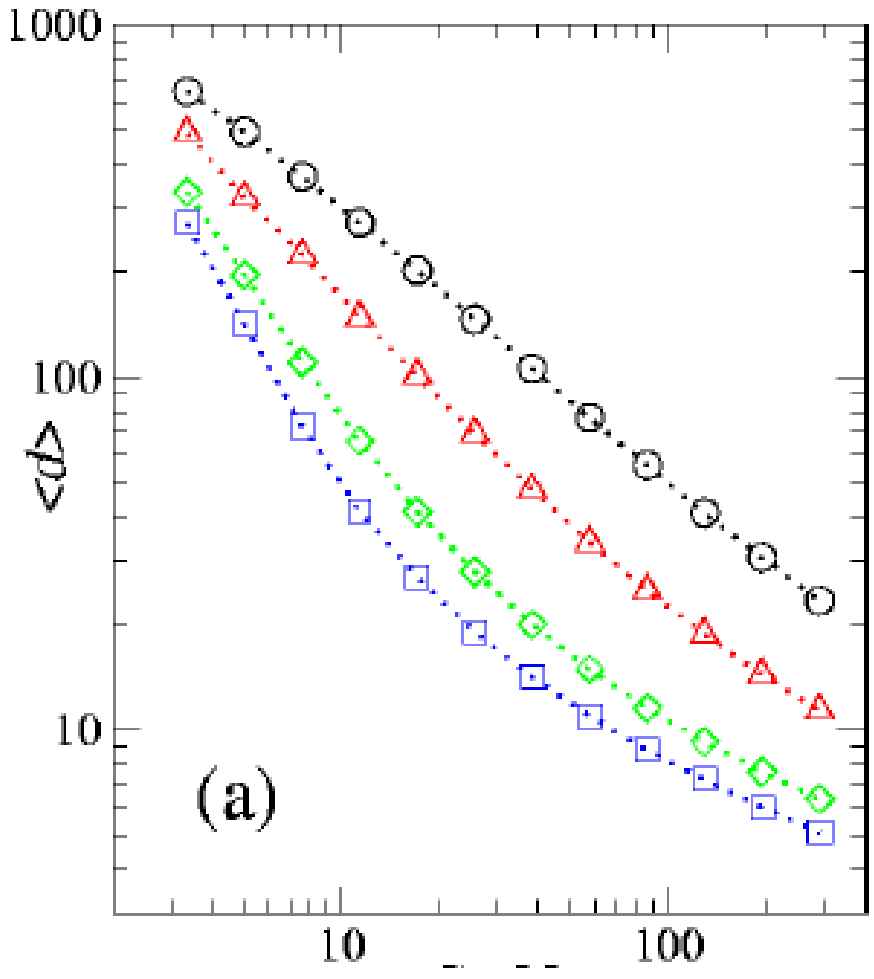
Average path length distribution over random graph ensemble





- Dispersion grows with growing α (at the same time the dispersion of the shortcut length distribution $P_s \sim s^{-\alpha}$ decrease with growing α)
  - numerical simulation is thanks to S.Polyakov (SINP MSU)
- $\Rightarrow$ the mean-field approximation is worse
- $\Rightarrow$ with growing α an importance of choice of specific sample of the network out of the hole ensemble grows

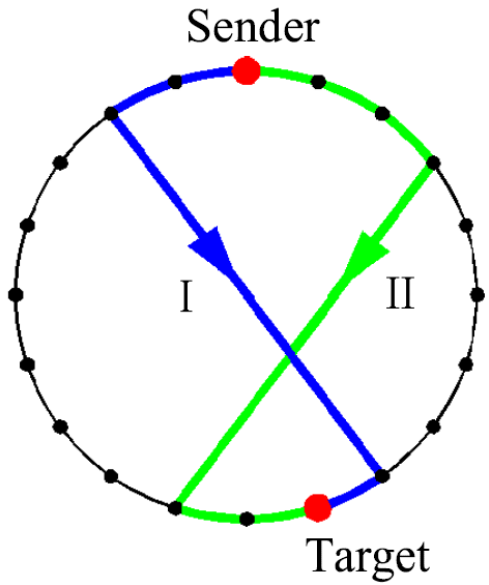
Optimization Problems for the Lattice Small-World



- A natural target functin: total wiring cost $C_W = pL^D\bar{s}$
- Numerical simulations: the mean distance decreases with growing α for a fixed value of C_W
  - The curves: black: $\alpha = 0$; red: $\alpha = 1$ green: $\alpha = 1.5$ blue: $\alpha = 1.75$;
- Analogously other network characteristics can be studied: link betweenness centrality, global and local efficiency, etc

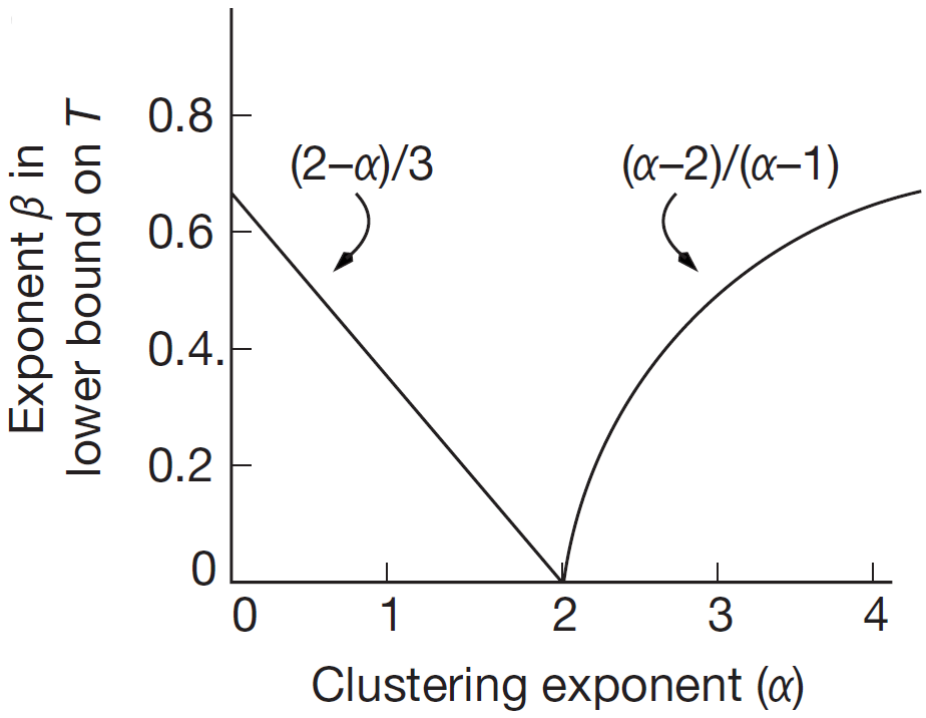
Local Navigation in Small-World Lattice Networks

- Pioneering works by Kleinberg (J. M. Kleinberg, Nature 406, 845 (2000)) initiated by the classical Milgram's experiment (1967)
- Local navigation in spatially embedded networks:
  - how a node can find a target (with known position in the basic lattice) efficiently with only a **local** knowledge of the network (the answer being trivial if you know the whole network)
- At one possible solution - greedy algorithm: at each step a message goes to a node closest to the target



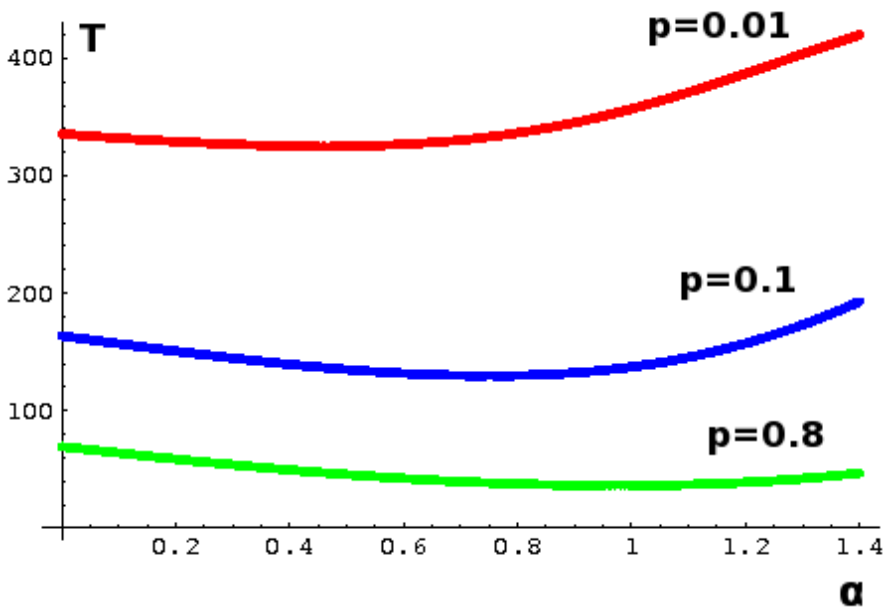
An example where global knowledge allows shorter path

Kleiberg's principal result



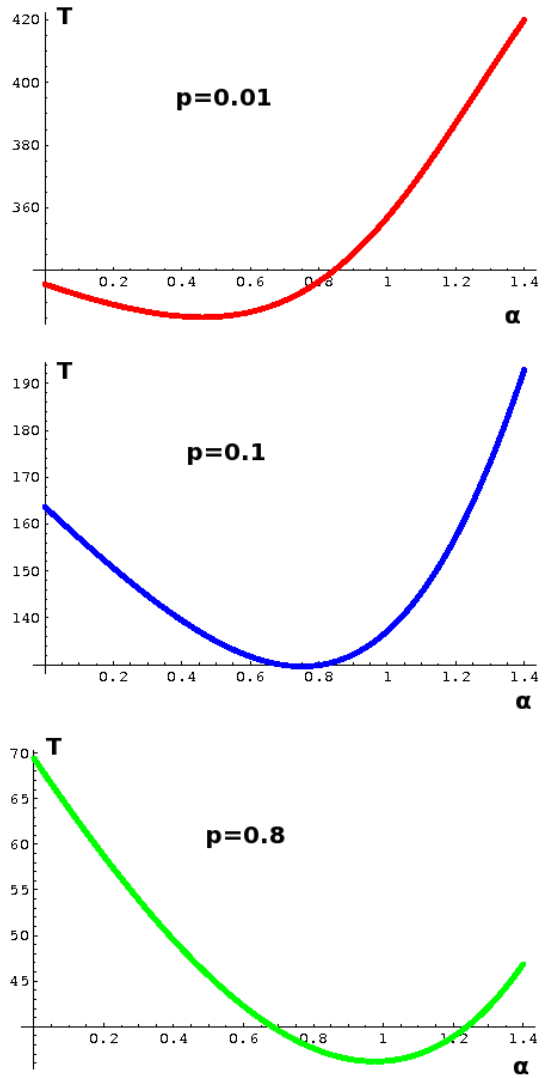
- Only $p = 1$ case was considered
- For $D = 2$ the plot shows lower bound for β in $\bar{T} \sim L^{\beta}$ $(N = L^2)$
- Most important: optimal value of shortcut length distribution is $\alpha = D$, which results in logarithmic dependence of the navigation time on the system size: $\bar{T} \sim \log^2 N$.

Generalization of Kleiberg's Result



- We have shown by using a bit simplified (unidirectional) local navigation algorithm and analitical calculations, that for $p \neq 1$ time minimum location $\alpha_{min} < D$ and depends on p
- For $D = 1$ the plot shows dependence the navigation time on α for three values of p

Generalization of Kleiberg's Result (2)



- For better visibility the plots from the previous transparency are shown separately
- It is seen that location of minima with respect to α depends on p

Conclusion

- Using standard regular topologies for design of interconnection networks for Exascale Supercomputers may result in inadmissible latencies
  - $\Rightarrow$ necessity to develop new approaches to the network design
- In this report we suggested to use irregular interconnection networks
  - small average path length between nodes is achieved by minimal addition of new links (shorcuts) to regular lattices
    - $\Rightarrow$ the optimal lattice structure for numerical simulations of $3D$-objects is preserved

Conclusion (2)

- Properties of such network are defined by two parameters
  - p - average density of the shortcuts
  - α - parameter of the shortcut length distribution $P_s \sim s^{-\alpha}$
- α defines the average shortcut length $\bar{s}$
  - thus $p\bar{s}$ is the average total shortcut length per compute node (the wiring cost)
  - from the point of view of the wiring cost large α is preferable (in the range $0 < \alpha < 2$)
- It is shown that choice of α for mininimization of local navigation time depends on p, the minimal time being achieved in the range $0 < \alpha \leq 1$
  - generalization of Kleiberg's Result for the case $p \neq 1$

Outlook

- It is necessary to develop principles of combined optimization of wiring cost (total shortcut length) and quality (latency)
- Investigations of flows of messages in interconnection small-world lattice networks
  - $\Rightarrow$ analytical models
  - $\Rightarrow$ numerical simulations
- Development of detailed algorithm(s) for creating optimal interconnection networks for exascale supercomputers in the class of lattice networks with random shortcuts
- Applications to the development of principles of communication environment design for neurocomputers and hybrid (bioelectronic) systems