UNIT V

Overview of simulator models and flow conditions. Methods of Solution. Performance Prediction. History match, concept on coning and compositional models. Stimulation Considerations.

OVERVIEW OF SIMULATOR MODELS AND FLOW CONDITIONS

Simulation of petroleum reservoir performance refers to the construction and operation of a model whose behavior assumes the appearance of actual reservoir behavior. The model itself is either physical (for example, a laboratory sandpack) or mathematical. A mathematical model is a set of equations that, subject to certain assumptions, describes the physical processes active in the reservoir. Although the model itself obviously lacks the reality of the reservoir, the behavior of a valid model simulates—assumes the appearance of—the actualreservoir.

The purpose of simulation is estimation of field performance (e.g., oil recovery) under one or more producing schemes. Whereas the field can be produced only once, at considerable expense, a model can be produced or run many times at low expense over a short period of time. Observation of model results that represent different producing conditions aids selection of an optimal set of producingconditionsforthereservoir.

The tools of reservoir simulation range from the intuition and judgment of the engineer to complex mathematical models requiring use of digital computers. The question is not whether to simulate, but rather which tool or method to use. This chapter concerns the numerical mathematical model requiring a digital computer. The Reservoir Simulation chapter in the 1987 edition of the *Petroleum Engineering Handbook* included a general description of reservoir simulation models, a discussion related to how and why they are used, choice of different types of models for different-reservoir problems, and reliability of simulation results in the face of model assumptions and uncertainty in reservoir-fluid and rock-description parameters. That material is largely omitted here. Instead, this chapter attempts to summarize current practices and trends related to development and application of reservoir simulation models.

Models have been referred to by type, such as black-oil, compositional, thermal, generalized, IMPES, Implicit, Sequential, Adaptive Implicit, or single-porosity, dual-porosity, and more. These types provide a confusing basis for discussing models; some refer to the application (e.g., thermal), others to the model formulation (e.g., implicit), and yet others to an attribute of the reservoir formation (e.g., dual-porosity). The historical trend, though irregular, has been and is toward the generalized model, which incorporates all the previously mentioned types and more. The generalized

model, which represents most models in use and under development today, will be discussed in this chapter. Current model capabilities, recent developments, and trends will then be discussed in relation to this generalized model.

(i)The Generalized Model

Any reservoir simulator consists of $n + m$ equations for each of N active gridblocks comprising the reservoir. These equations represent conservation of mass of each of n components in each gridblock over a timestep Δt from t^n to t^{n+1}. The first n (primary) equations simply express conservation of mass for each of n components such as oil, gas, methane, CO_2, and water, denoted by subscript $I = 1,2,\ldots, n$. In the thermal case, one of the "components" is energy and its equation expresses conservation of energy. An additional m (secondary or constraint) equations express constraints such as equal fugacities of each component in all phases where it is present, and the volume balance $S_w + S_o + S_g + S_{solid} = 1.0$, where S solid represents any immobile phase such as precipitated solid salt or coke.

There must be $n + m$ variables (unknowns) corresponding to these $n + m$ equations. For example, consider the isothermal, three-phase, compositional case with all components present in all three phases. There are $m = 2n + 1$ constraint equations consisting of the volume balance and the $2n$ equations expressing equal fugacities of each component in all three phases, for a total of $n + m = 3n + 1$ equations. There are $3n + 1$ unknowns: p, S_w, S_o, S_g, and the $3(n - 1)$ independent mol fractions x_{ij}, where $i = 1,2,\ldots, n - 1$; $j = 1,2,3$ denotes the three phases oil, gas, and water. For other cases, such as thermal, dual-porosity, and so on, the m constraint equations, the n + m variables, and equal numbers of equations and unknowns can be defined for each gridblock.

Because the m constraint equations for a block involve unknowns only in the given block, they can be used to eliminate the $m$secondary variables from the block's n primary or conservation equations. Thus, in each block, only n primary equations in n unknowns need be considered in discussions of model formulation and the linear solver. The n unknowns are denoted by P_{i1}, P_{i2} ,..., P_{in}, where $P_{in}$is chosen as pressure p_i with no loss of generality. These primary variables may be chosen as any n independent variables from the many available variables: phase and overall mol fractions, mol numbers, saturations, p, and so on. Different authors choose different variables.[12][13][14][15] Any sensible choice of variables and ordering of the primary equations gives for each gridblock a set of $n$equations in n unknowns which is susceptible to normal Gaussian elimination without pivoting. The (Newton-Raphson) convergence rate for the model's timestep calculation is independent of the variable choice; the model speed (CPU time) is essentially independent of variable choice.

The Ith primary or conservation equation for block i is

$$M_{iI}^{n+1} - M_{iI}^{n} = \Delta t\left(\sum_{j=1}^{j=N} q_{ijI} - q_{iI}\right) I = 1,2,...n, \quad \text{.....(1)}$$

where M_{iI} is mass of component I in gridblock i, q_{ijI} is the interblock flow rate of component I from neighbor block j to block i, and q_{iI} is a well term. With transposition, this equation is represented by $f_{iI} = 0$, the Ith equation of gridblock i. All n equations $f_{iI} = 0$ for the block can be expressed as the vector equation $\boldsymbol{F}_i = 0$ where f_{iI} is the Ith element of the vector F_i. Finally, the vector equation

$$\mathbf{F}(\mathbf{P}_1, \mathbf{P}_2, ..., \mathbf{P}_N) = 0 \quad \text{.....(2)}$$

represents the entire model, where the ith element of the vector $\mathbf{F}$ is $\mathbf{F}_i$. $\mathbf{F}$ is a function of the N vector unknowns $\mathbf{P}_i$, where the Ith scalar element of $\mathbf{P}_i$ is P_{iI}. Application of the Newton-Raphson method gives

$$\mathbf{F}^l + \delta\mathbf{F} = \mathbf{F}^l + A\delta\mathbf{P} = 0, \quad \text{.....(3)}$$

where $\delta\mathbf{P}$ is $\mathbf{P}^{l+1} - \mathbf{P}^l$ and the $N \times N$ matrix A represents the Jacobian $\partial\mathbf{F}/\partial\mathbf{P}$. The element A_{ij} of A is itself an $n \times n$ matrix $\partial\mathbf{F}_i/\partial\mathbf{P}_j$ with scalar elements $a_{rs} = \partial f_{ir}/\partial P_{js}$, r and s each = 1,2,..., n. **Eq. 3** is solved by the model's linear solver. The matrix A is very sparse because A_{ij} is 0 unless block j is a neighbor of block i.

The calculations for a timestep consist of a number of Newton (*nonlinear* or *outer*) iterations terminated by satisfaction of specified convergence criteria. Each Newton iteration requires:

(a) Linearization of the constraint equations and conservation **Eq. .1**.
(b) Linear algebra to generate the A matrix coefficients.
(c) Iterative solution of **Eq. 3** (*inner* or *linear* iterations).
(d) Use of the new iterate $\mathbf{P}^{l+1}$ to obtain from **Eq. 1** the moles of each component in the gridblock.
(e) A flash to give phase compositions, densities, and saturations which allow generation of the A matrix coefficients for the next Newton iteration.

(ii) Model Formulations

A major portion of the model's total CPU time is often spent in the linear solver solution of **Eq. 3**. This CPU time in turn reflects the many multiply operations required. The model formulation has a large effect on the nature and expense of those multiplies.

Implicit vs. Explicit. The interblock flow term in **Eq. 1**,

$$q_{ijI} = T_{ij} \sum_{J=1}^{J=3} \lambda_j \rho_J x_{IJ} \left(\Delta p_J - \gamma_J \Delta Z \right), \quad \text{.....(4)}$$

uses phase mobilities, densities, and mol fractions evaluated at the upstream blocks. A gridblock is *implicit* in, say, the variable S_g if the new time level value S_g^{n+1} is used to evaluate interblock flow terms dependent upon it. The block is *explicit* in S_g if the old time level value S_g^n is used.

The Implicit Forumulation. The implicit formulation[16] expresses interblock flow terms using implicit (new time level) values of all variables in all gridblocks. As a consequence, all nonzero A_{ij} elements of the A matrix of **Eq. 3** are full $n \times n$ matrices. The resulting multiplies in the linear solver are then either matrix-matrix or matrix-vector multiplies, requiring work (number of scalar multiplies) of order n^3 or n^2, respectively.

The IMPES Formulation. Early paper[17][18][19] presented the basis of the IMPES (implicit pressure, explicit saturations) formulation for the black-oil case: take all variables in the interblock flow terms explicit, except for pressure, and eliminate all nonpressure variables from the linearized expressions for M_{il}^{n+1} in **Eq. 1**. The obvious extension to any type model with any number of components was presented later,[20] and numerous IMPES-type compositional models have been published.

The model **Eq. 3** can be written as:

$$A_{ii} \delta \mathbf{P}_i + \sum_{j \neq i} A_{ij} \delta \mathbf{P}_j = -\mathbf{F}_i^l \; i = 1,2,\ldots, N \; . \quad \text{.....(5)}$$

If all variables but pressure are explicit in the interblock flow terms, then all entries but those in the last column of the $n \times n$ A_{ij} ($j \neq i$) matrix are zero (recall, the n th variable in each gridblock, P_{in}, is pressure p_i). This allows elimination of all nonpressure variables and reduction of the vector **Eq. 5** to the scalar equation in pressure only

$$a_{ii}\delta p_i + \sum_{j \neq i} a_{ij}\delta p_j = -f_i^l \; i = 1,2,\ldots, N \quad \ldots\ldots\ldots\ldots\ldots\ldots(6)$$

or

$$A\delta\mathbf{P} = -\mathbf{F}^l, \quad \ldots\ldots\ldots\ldots\ldots\ldots(7)$$

where A is now a scalar $N \times N$ matrix and the **P** and **F** vectors have N scalar elements p_i and f_i, respectively. The multiplications required in solution of the IMPES pressure **Eq. 7** are scalar multiplications, requiring a small fraction of the work of the matrix-matrix and matrix-vector multiplications of the implicit formulation. Thus, the model CPU time per gridblock per Newton iteration for moderate or large n is much less for the IMPES formulation than for the implicit formulation.

The Sequential Formulation. The stability of the IMPES formulation for the two-phase water/oil case was improved by following the IMPES pressure equation solution with solution of a water saturation equation using implicit saturations (mobilities).[23] This concept was extended to the three-phase case and called the *sequential* formulation.[24] For each Newton iteration, this method requires solution of the IMPES pressure **Eq. 7**, followed by solution for two saturations from a similar equation where the A_{ij} elements of A are 2 × 2 matrices.

A sequential compositional model was described[15] and mentioned the desirability of a sequential implicit treatment of mol fractions in addition to saturations.

The Adaptive Implicit Forumlation. The Adaptive Implicit Method (AIM) [25] uses different levels of implicitness in different blocks. In each gridblock, each of the n variables may be chosen explicit or implicit, independent of the choices in other gridblocks. The choices may change from one timestep to the next. This results in the same equation $A\delta\mathbf{P} = -\mathbf{F}^l$ as the Implicit formulation except that the elements A_{ij} of the A matrix are rectangular matrices of variable size. The numbers of rows and columns in A_{ij} equal the numbers of implicit variables in blocks i and j, respectively; all A_{ii} are square matrices. The CPU expense per Newton iteration of an AIM model lies between those of IMPES and Implicit models, tending toward the former as more blocks are taken implicit in pressureonly.

Choice of Formulation. For a given problem, the previous four formulations generally give widely different CPU times. Generalizations regarding the best formulation have many exceptions. Arguably, the trend is or should be toward sole use of the AIM formulation. This is discussed in the Stable Step and Switching Criteria sections to follow. Current simulation studies use all of these

formulations. The Implicit formulation is generally faster than IMPES for single-well coning studies, and for thermal and naturally fractured reservoir problems. For other problems, IMPES is generally faster than Implicit for moderate or large n (say, $n > 4$). Most participants used IMPES for SPE Comparative Solution Project problems SPE1, SPE3, SPE5, and SPE10. All participants used the Implicit formulation for SPE2, SPE4, SPE6, and SPE9. No participants in SPE1 through SPE10 used a Sequential model, and, with few exceptions, none used AIM.

A frequently stated generalization is that numerical dispersion error is significantly larger for Implicit than for IMPES formulations. Truncation error analysis shows this error to be proportional to $\Delta x + u\Delta t$ for Implicit and $\Delta x - u\Delta t$ for IMPES. Real problem nonlinearities and heterogeneity render the analysis approximate and the generalization of limited merit.

(iii)Advances in Model Forumlations

The IMPES formulation was improved by concepts of relaxed volume,[13][14][15] better choice of variables, 13 and "adaptive" flash calculations.[13]

Relaxed Volume. The relaxed volume concept relates to the timestep calculation Steps (d) and (e) given previously. Step (d) gives the mass of each component in the gridblock, M_I^{l+1}, which in turn gives overall composition $\{z_I\}^{l+1}$. The Step (e) flash then gives phase amounts and densities which in turn give new iterate S_w, S_o, and S_g values. These saturations do not sum to 1.0 because of the nonlinear nature of the conservation **Eq. 17.1**. If the saturations are altered (e.g., divide each by their sum) to exactly satisfy the volume balance $\Sigma_J S_J = 1$, then an incremental (timestep) mass-balance error occurs. If the saturations are not altered, then mass is conserved but there is a volume-balance error $\Sigma_J S_J - 1$. The authors[13][14][15] chose to preserve mass and *carry forward* the volume balance error from iterate to iterate and step to step. The volume balance going into iterate $l + 1$ is $\Sigma_J \delta S_J = 1 - \Sigma_J S_J^l$. This in effect conserves both mass and volume because there is no permanent or accumulating volume error—only that of the given timestep. Equally important, there is no need to iterate out the volume error to a "tight" tolerance, and Newton iterations and model CPU are reduced. In contrast, the previous or historical IMPES procedure reset saturations to preserve volume and iterated out the mass-balance error. Because the latter error was not carried forward, more Newton iteration (and CPU time) was required to keep the permanent, accumulating mass balance error tolerably low. This use of relaxed volume with carryover also reduces Newton iterations and CPU time in the Implicit formulation.[21]

This discussion implies some fundamental advantage of preserving mass and iterating out volume error as opposed to preserving volume and iterating out mass error. In the writer's opinion, that is not true provided the error is carried forward in both cases. The Newton iteration requirement and CPU time should be similar if "equivalent" mass and volume error tolerances are used as convergence criteria.

Variable Choice. The linear algebra required to reduce the gridblock's n conservation equations to the IMPES pressure equation is influenced by the choice of variables. The influence is absent for black oil, moderate for "moderate" n and up to a factor of three for large n (say, > 15).[1] The choices of p and mol fractions $\{z_I$ or mol numbers[14][15] are better than the choice of p, saturations, and phase mol fractions[12] for large n. The effect of this variable choice on total CPU time is often small because the affected work is often a small part of total CPU time. This IMPES reduction is absent in the Implicit formulation and the last of the above variable choices is arguably preferable.[22]

Adaptive Flash Calculations.[13] The work of EOS flash calculations, including the generation of fugacities and their derivatives, can significantly affect model efficiency when the linear solver does not dominate total CPU time. There may be little need to perform (most of) that work in a gridblock when p and composition are changing slowly. Use of internal, intelligent criteria dictating when that work is needed can significantly reduce the total-run flash calculation CPU time.[13] This is similar in principle to the AIM selection of explicit variables for gridblocks which are quiescent in respect to throughput ratio.

Stable Timestep and Switching Criteria

This topic relates to the observation that lower run turnaround time can increase benefits from a reservoir study allotted a budgeted time period. As a corollary, time spent in repeated runs fighting model instabilities or time-stepping is counterproductive. While many factors affect this run time, it always equals the product (CPU time/step) × (number of timesteps). The first factor is "large" and the second "small" for the Implicit formulation, and conversely for the IMPES formulation. IMPES is a conditionally stable formulation requiring that $\Delta t < \Delta t^*$ to prevent oscillations and error growth, where Δt^* is maximum stable timestep. The conditional stability stems from the explicit treatment of nonpressure variables in the interblock flow terms. Mathematicians performed stability analyses for constant-coefficient difference equations bearing some resemblance to IMPES. Authors in our industry extended and applied their results to derive expressions for Δt^*, in particular,[27]

$$\Delta t^* = \frac{V_p}{f_g'\left(|q_x| + |q_y| + |q_z|\right) + 2P_{cgo}'\psi\left(T_x + T_y + T_z\right)} \quad \text{.....(8)}$$

for the black-oil 3D case of gas/oil flow. This shows that stable step Δt^* is dependent upon flow rates, phase mobility, and capillary pressure derivatives, which of course vary with time and from one gridblock to another. Thus, at a given timestep, there are block-dependent stable step values Δt^*_i, where $1 < i < N$, and the IMPES stable step is Min(i) Δt^*_i. An IMPES model using this internally determined stable step will

run stably but may suffer from the weakest-link principle. As an extreme example, consider a 500,000-gridblock problem where, over a 100-day period, the Δt^*_i value is 0.01 day for one block and > 30 days for all other blocks. The IMPES model will require 10,000 timesteps over the 100-day period.

In the AIM formulation, the stable step Δt^*_i depends upon the number and identities of variables chosen explicit in block i; theoretically, $\Delta t^*_i = \infty$ if all block i variables are chosen implicit. In the previous example, all nonpressure variables could be chosen implicit in the block where $\Delta t^*_i = 0.01$ and explicit in all other blocks. The AIM model would then require CPU time/step essentially no greater than the IMPES model but would require only three timesteps for the 100-day period.

Numerous papers[28][29][30][31][32][33] address the problem of determining expressions for the Δt^*_i for use internally as switching criteria to select block variables as explicit or implicit in the AIM model. The stability analyses involved are complex and may be impractically complex when allowing the implicit vs. explicit variable choice to include all permutations (in number and identity) of the n variables. The most reliable and efficient AIM models in the future will stem from continuing research leading to the following: (a) Δt^*_i estimates which are "accurate," and (b) implicit vs. explicit variable choices, block by block, which are near-optimal[34] and minimize total CPU time, (CPU time/step) × (number of steps).

(iv)The Linear Solver

Preconditioned Orthomi is the most widely used method for iterative solution of **Eqs. 3** or**7**. Nested Factorization (NF) [36] and incomplete LU factorization [ILU(n)] [37] are the two most widely used preconditioners. The term "LU factorization" refers to the factoring of the matrix A into the product of a lower triangular matrix L and an upper triangular matrix U. That is an expensive operation but is straightforward, involving only Gaussian elimination. The term "ILU(n)" denotes incomplete LU factorization, where only limited fill-in is allowed and n is the "order of fill."[37] NF performs exceptionally well when transmissibilities associated with a particular direction (in a structured grid) dominate those in other directions uniformly throughout the grid. In general, ILU(n) or red-black ILU(n)[38] [RBILU(n)] is less sensitive than NF to ordering of the blocks and spatial variation of the direction of dominant transmissibilities. In addition, RBILU(n) or ILU(n) have the parameter n (order of allowed infill) which can be increased as needed to solve problems of any difficulty.

A literature search and discussions with numerous developers and users have failed to establish consensus on whether NF or ILU preconditioning is better. Some are strong advocates of one method and others are just as adamantly supportive of the other. But many find, like this writer, that the better method is problem-dependent and it is difficult to find a reliable *a priori* indicator for making an up-front choice. In the writer's experience, (a) when NF works well, it is faster than ILU methods, (b) RBILU(0) with no residual constraint is frequently the best

of the ILU variants and a good default choice, and (c) in some cases, global residual constraint with the ILU or RBILU method is beneficial.

(v)Cartesian Grids and Reservoir Definition

For many years, simulation used orthogonal Cartesian grids. In the past 15 years, numerous papers have described local grid refinement and various non-Cartesian grids, as discussed in the Gridding section. These papers show that non-Cartesian grids can reduce grid-orientation effects and provide definition and accuracy near wells, faults, highly heterogeneous areas, and so on more efficiently than Cartesian grids. The premise that Cartesian grids cannot provide required accuracy efficiently in these respects has come to be accepted as a fact. In addition, advances in geophysics have led to geostatistical description of permeability and porosity on a fine scale once unimaginable. Increasingly, our papers include examples using thousands of gridblocks for two- or few-well "patterns," in part to reflect these geostatistical descriptions. The purpose of this section is to show, using a few examples, that Cartesian grids can provide adequate accuracy and reservoir and near-well definition efficiently in some cases, even without local grid refinement. No generalizations from the examples used are intended. For the most part, the examples are taken from the literature.

METHODS OF SOLUTION

Preconditioned Orthomin is the most widely used method for iterative solution Nested Factorization (NF) and incomplete LU factorization [ILU(*n*)] are the two most widely used pre conditioners.

The term "LU factorization" refers to the factoring of the matrix A into the product of a lower triangular matrix L and an upper triangular matrix U. That is an expensive operation but is straightforward, involving only Gaussian elimination.

The term "ILU(*n*)" denotes incomplete LU factorization, where only limited fill-in is allowed and *n* is the "order of fill."[37]

NF performs exceptionally well when transmissibilities associated with a particular direction (in a structured grid) dominate those in other directions uniformly throughout the grid. In general, ILU(*n*) or red-black ILU(*n*)[38] [RBILU(*n*)] is less sensitive than NF to ordering of the blocks and spatial variation of the direction of dominant transmissibilities. In addition, RBILU(*n*) or ILU(*n*) have the parameter *n* (order of allowed infill) which can be increased as needed to solve problems of any difficulty.

A literature search and discussions with numerous developers and users have failed to establish consensus on whether NF or ILU preconditioning is better. Some are strong advocates of one method and others are just as adamantly supportive of the other. But many find, like this writer, that the better method is problem-dependent and it is difficult to find a reliable *a priori* indicator for making an up-

front choice. In the writer's experience, (a) when NF works well, it is faster than ILU methods, (b) RBILU(0) with no residual constraint is frequently the best of the ILU variants and a good default choice, and (c) in some cases, global residual constraint with the ILU or RBILU method is beneficial.

PERFORMANCE PREDICTION

The reservoir simulation model-building process and history matching are intended to provide a working model of the reservoir and establish a level of confidence in the validity of a flow model. Therefore, the final history matched model is usually re-configured to predict the behavior of the reservoir into the future. When a reservoir simulation model is changed from history matching to prediction mode, the phase rate profiles should be smooth, provided new wells are not added or existing wells shut-in, and the fundamental constraints on the wells are not changed. There should not be a shift up or down in rates at this point. Such a shift is usually indicative of non-calibrated wells.

It is recommended that the last year of history is run in prediction mode and the actual production compared with the simulated prediction. While this should not be expected to give a perfect match, it will help to highlight major discrepancies in the model.When a reservoir simulation model is used for predictions, the limitations and uncertainties involved in the reservoir simulation models should be recognized. If the geological model, for example, is not reasonable and observed data quality is poor, not much quality can be expected from reservoir simulation model, no matter the quality of the history match.

HISTORY MATCHING

The main objective of the history match is to achieve a reasonable agreement between the simulated and observed historical field/well behavior to establish a satisfactory quality reservoir management tool. This is done under the premises that the geological model, the reservoir parameters, and other static and dynamic data used have a "defendable" quality.

(i)Manual vs. assisted history matching

Two approaches can be applied for performing history matching study: manual history matching and assisted history matching using specialized software. Traditionally, history matching is performed by a trial-and-error approach. In this case, a lot of manual tasks are involved, such as changing the reservoir simulation model, running reservoir simulations, plotting curves and comparing to observed data. The main advantage of assisted history matching is to automate those manual tasks, such as reservoir simulation model modifications, running reservoir simulations, comparison of observed and reservoir simulation data, etc. however care should be taken in setting parameter range limits, etc. in automated history match to ensure any solutions are physically valid.

(ii)History matching input data

The following historical (measured) input data for individual wells or reservoirs are typically used in history matching process:

- RFT pressures (measured pressure points vs. depth)
- Shut-in pressure (measured pressure vs time)
- Historical production / injection rates vs. time
- Allocated or measured well GOR and WCT vs. time
- Fluid saturation profiles from well logs

(iii)History matching steps

The following steps are recommended for performing history matching:

- Match average reservoir pressure and field rates to have a good understanding about material balance in the reservoir.
- Match individual well RFT pressure to have control on compartmentalization and flow barriers.
- Match individual well gas/oil ratio, water-cut and shut-in pressure to have a good control on flow dynamics in reservoir and well performance.

(iv)History match quality

There are several ways to decide if a match is satisfactory. In all cases, a clear understanding of the study objectives should be the reference for making the decisions. For example, if a coarse study is being performed, the quality of the match between observed and simulated parameters does not need to be as accurate as it would be for a more detailed study.

Quality of Modifications Made. If the model has a good match but the changes made were not realistic, then the model results should be viewed with skepticism. Remember that the ultimate objective of reservoir simulation is not achieving a history match; it is being able to reasonably predict the future performance of the reservoir. The history match is only an intermediate step in the modeling process.

CONCEPT OF CONING

Coning is a production problem in which gas cap gas or bottom water infiltrates the perforation zone in the near-wellbore area and reduces oil production. Gas coning is distinctly different from, and should not be confused with, free-gas production caused by a naturally expanding gas cap. Likewise, water coning should not be confused with water production caused by a rising water/oil contact (WOC) from water influx. Coning is a rate-sensitive phenomenon generally associated with high producing rates. Strictly a near-wellbore phenomenon, it only develops once the pressure forces drawing fluids toward the

wellbore overcome the natural buoyancy forces that segregate gas and water from oil.

(i)Basis of terminology

The term coning is used because, in a vertical well, the shape of the interface when a well is producing the second fluid resembles an upright or inverted cone (**Fig. 1**). Important examples of coning include:

- Production of water in an oil well with bottomwater drive
- Production of gas in an oil well overlain by a gas cap
- Production of bottom water in a gas well

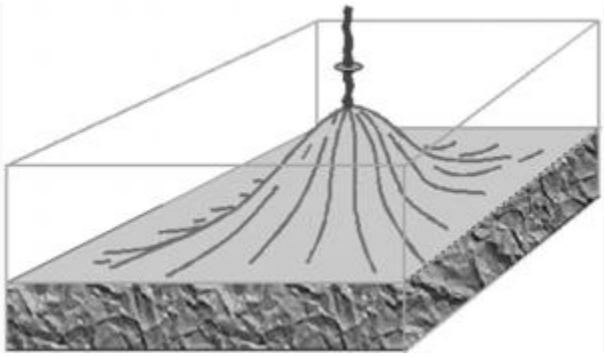

Fig. 1 – Coning in a vertical well.

In a horizontal well, the cone becomes more of a crest (**Fig. 2**), but the phenomenon is still customarily called coning. In a given reservoir, the amount of undesired second fluid a horizontal well produces is usually less than for a vertical well under comparable conditions. This is a major motivation for drilling horizontal wells, for example, in thin oil columns underlain by water.

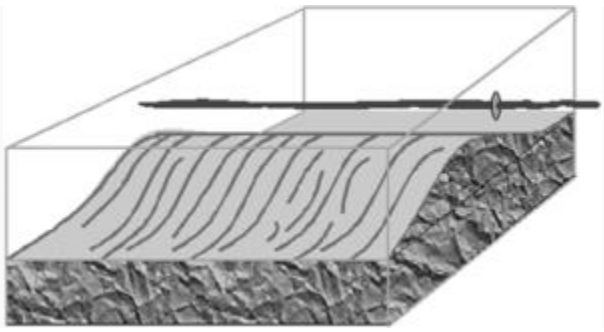

Fig. 2 – Coning in a horizontal well.

(ii)Impact of coning

Coning is a problem because the second phase must be handled at the surface in addition to the desired hydrocarbon phase, and the production rate of the

hydrocarbon flow is usually dramatically reduced after the cone breaks through into the producing well. Produced water must also be disposed of. Gas produced from coning in an oil well may have a market, but also may not. In any event, production of gas in an oil well after the cone breaks through can rapidly deplete reservoir pressure and, for that reason, may force shut in of the oil well.

Several strategies may apply to wells with a potential to cone. One is to try to predict the rate at which a well will cone and produce at a lower rate as long as possible. Or, optimal economics may result by producing at a much higher rate, causing the well to cone, but increasing the cumulative hydrocarbon volume produced (and present value) at any future date. A horizontal well may be preferred to a vertical well.

(iii)Predicting coning

Most prediction methods for coning predict a "critical rate" at which a stable cone can exist from the fluid contact to the nearest perforations. The theory is that, at rates below the critical rate, the cone will not reach the perforations and the well will produce the desired single phase. At rates equal to or greater than the critical rate, the second fluid will eventually be produced and will increase in amount with time. However, these theories based on critical rates do not predict when breakthrough will occur nor do they predict water/oil ratio or gas/oil ratio (GOR) after breakthrough. Other theories predict these time behaviors, but their accuracy is limited because of simplifying assumptions.

The calculated critical rate is valid only for a certain fixed distance between the fluid contact and the perforations. With time, that distance usually decreases (for example, bottom water will usually tend to rise toward the perforations). Thus, the critical rate will tend to decrease with time, and the economics of a well with a tendency to cone will continue to deteriorate with time.

Whether a cone will move toward perforations depends on the relative significance of viscous and gravitational forces near a well. The pressure drawdown at the perforations tends to cause the undesired fluid to move toward the perforations. Gravitational forces tend to cause the undesired fluid to stay away from the perforations. Coning occurs when viscous forces dominate.

The variables that could affect coning are:

- Density differences between water and oil, gas and oil, or gas and water (gravitational forces)
- Fluid viscosities and relative permeabilities
- Vertical and horizontal permeabilities
- Distances from contacts to perforations.

Coning tendency turns out to be quite dependent on some of these variables and insensitive to others.

A number of prediction methods have been published. There is no guarantee of great accuracy when using any of these methods because they all contain significant simplifying assumptions. In particular, areal and vertical variations in vertical permeability (because of flow barriers of varying extent) can cause the prediction methods to differ significantly from what actually happens in the field. Accordingly, the prediction methods are best used for quick approximations, screening, and comparison of alternatives. Reservoir simulations, based on accurate reservoir characterization, will ultimately be required.

The coning prediction method proposed by Chaperon[1] is of particular interest because of the variables it includes and because variations of the method are applicable to gas and water coning in both vertical and horizontal wells. For vertical wells, the Chaperon method calculates the critical rate for coning from the expression

$$q_c = \frac{4.888 \times 10^{-4} k_h h_c^2 \Delta\rho q_{cD}}{B_o \mu_o}, \quad \text{.....(1)}$$

where

$$q_{cD} = 0.7311 + \frac{1.843 / r_{rD}}{\sqrt{k_v / k_h}}, \quad \text{.....(2)}$$

$$r_{rD} = \left(r_e / h_c\right)\sqrt{k_v / k_h}, \quad \text{.....(3)}$$

$$\Delta\rho = \rho_w - \rho_o \text{ or } \rho_o - \rho_g, \text{ density difference, } \mathrm{g/cm^3}, \quad \text{.....(4)}$$

and h_c = distance from perforations to fluid contact, ft. For horizontal wells, the critical rate is given by

$$q_c = \frac{4.888 \times 10^{-4} L_w}{\left(a_H / 2\right)} \frac{\Delta\rho\left(k_h h^2\right)}{\mu_o B_o} F, \quad \text{.....(5)}$$

where

$$F = 3.964 + 0.0616 a_{HD} - 0.000540 a_{HD}^2 \quad \text{.....(6)}$$

and

$$a_{HD} = \left(\frac{a_H}{2h_c}\right)\sqrt{\frac{k_v}{k_h}} \quad \text{.....(7)}$$

(iv)Coning strategies

Under ideal conditions in which no coning exists, flow is principally horizontal and mainly oil is produced. **Fig. 3** illustrates a producing well with no coning. When coning exists, however, the overlying gas is drawn downward or bottomwater is drawn upward and into the oil column. Coning trades oil production for gas or water production. **Fig. 4** illustrates a producing well subject to gas and water coning.

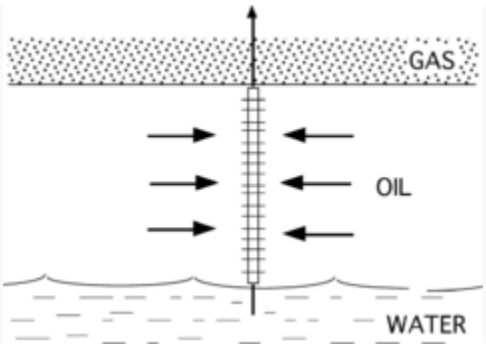

Fig. 3 – A producing well with no coning.

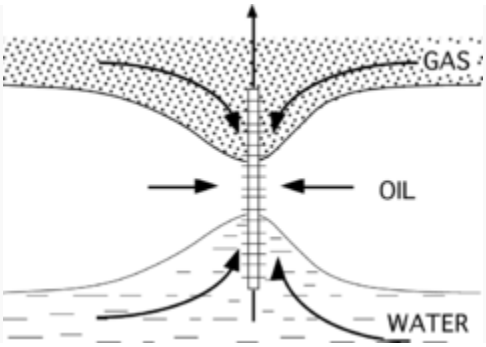

Fig. 4 – A producing well subject to gas and water coning.

Two strategies commonly are used to minimize coning. One approach is partial perforation or penetration. In this approach, only a limited portion of the pay thickness is perforated. If gas coning is anticipated, the pay thickness near the GOC is not perforated. If water coning is anticipated, the pay thickness near the WOC is not perforated. In instances in which severe coning is expected, only a small portion of the pay thickness may be perforated. The variables in **Fig. 5** define the length of the perforation interval, b, and its position within the pay thickness, h. The distance L_g is the distance between the top of the pay and the uppermost perforation, and the distance L_w is the distance between the bottom of the pay and the lowest perforation. The quotient b/h is the partial perforation fraction. Although this strategy will reduce and can eliminate coning problems, it suffers an obvious drawback; namely, it temporarily reduces oil production in the hope of eventually avoiding coning.

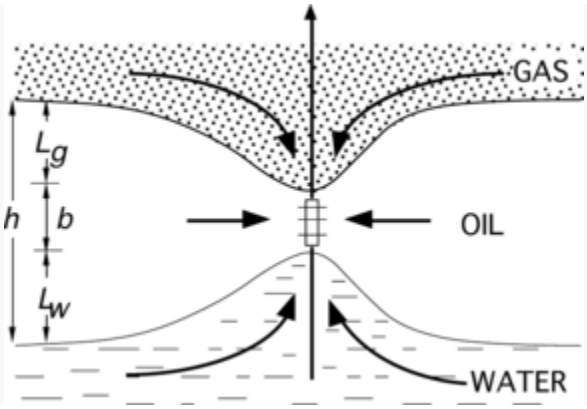

Fig. 5 – Definition of variables for a partially perforated producing well.

A second remedial strategy is based on the observation that there is a critical producing rate below which the cone stabilizes and will not reach the perforations. This critical rate is a function of the perforation length. As the perforation length increases, the critical producing rate decreases. Often, the critical producing rate is much less than the possible producing rate. This difference creates an operational decision:

- Produce at a rate greater than the critical and eventually risk coning
- Produce at a rate less than the critical and temporarily sacrifice oil production

If the critical rate is less than the minimum economic rate, then the operator has no choice but to produce above the critical rate or abandon the well.

To combat coning, a hybrid strategy is often used whereby a combination of partial perforation and a reduced producing rate is used. One especially unattractive consequence of gas coning is that it prematurely depletes the gas-cap gas and diminishes the gas-cap producing mechanism. Fortunately, gas

coning is not as problematic as water coning because the density difference between oil and gas is greater than the difference between water and oil. This density difference through gravity segregation helps mitigate coning.

To develop an effective remedial strategy against coning, certain theoretical aspects regarding coning must be understood. Mathematically, coning is a challenging problem because of its complexity. To develop tractable analytical solutions, tenuous assumptions must be invoked. These assumptions limit the practical applicability of these solutions. The most reliable way to study coning is with a specially designed finite-difference simulator. Nevertheless, certain analytical solutions and empirical correlations can be helpful and serve as a preliminary guide.

Muskat and Wyckoff and Chaney *et al.*were among the first to contribute substantively to this problem. Since their efforts, several other authors have contributed to the body of literature. Many of these works have led to similar correlations. Wheatleypresented a comparison of some popular correlations. As a representative sample, the correlations of Schols and Chierici *et al.* are presented here. Both works apply to both water and gas coning. Both efforts also use the following equation to compute the critical producing rate:

$$q_c = \frac{0.003073 k_o h^2 \Delta\rho}{\mu_o B_o} q_{Dc} \quad \text{.....(8)}$$

where:

- $\Delta\rho$ = density difference (g/cm^3)
- B_o = average oil formation volume factor (FVF)
- μ_o = average oil viscosity (cp)
- k_o = oil permeability (md)
- q_{Dc} = dimensionless critical producing rate
- h = pay thickness (ft)
- q_c is given in STB/D

The oil permeability, k_o, is the product of the horizontal permeability and the oil relative permeability. The dimensionless critical rate, q_{Dc}, is specified by correlation.

(v)Variables affecting coning

The ratio of q_c/q is a measure of the tendency not to cone. As q c increases or q decreases, the likeliness to avoid increases.

$$\frac{q_c}{q} \propto \frac{h^2 \Delta\rho \left[\ln\left(\frac{r_e}{r_w}\right) + s \right]}{b} \quad \text{.....(9)}$$

This expression shows that the likeliness to control coning increases as the penetration interval *b* decreases. **Eq. 9** also shows that the likeliness to control coning increases as the pay thickness increases, density difference increases, well spacing increases, and perforation length decreases. Horizontal permeability does not affect the likelihood of success. This expression also suggests that controlling coning in a thin reservoir may be difficult.

(vi)Additional measures to control coning

Other techniques have been applied to control coning. These include:

- Placing an artificial barrier above or below the pay to suppress vertical flow
- Injecting oil to control gas coning
- Use of horizontal wells

Barriers composed of cement and high-molecular-weight polymers have been tried. Another, although expensive, technique is to drill additional wells and produce them at the critical rate.

COMPOSITIONAL MODELS

Prediction of a miscible flood is best done with a compositional reservoir simulator. The simulation must be able to predict the phase behavior as well as the sweep behavior in the reservoir to forecast such quantities as incremental oil recovery, miscible-solvent requirement, and solvent utilization efficiency and to optimize such variables as solvent composition, operating pressure, slug size, water-alternating-gas (WAG) ratio, injection-well placement, and injection rate.

The compositional reservoir simulator calculates the flow in up to three dimensions of solvent, oil, and water phases as well as *n*components in the solvent and oil phases. It also computes the phase equilibrium of the oil and solvent phases (i.e., the equilibrium compositions and relative volumes of the solvent and oil phases) in each gridblock of the simulator. In addition, it computes solvent- and oil-phase densities. The equilibrium compositions and densities are calculated with an equation of state (EOS). From knowledge of the phase compositions and densities, solvent and oil viscosity and other properties such as interfacial tension are estimated from correlations.

(i)Predicting phase behavior

Phase behavior can be predicted by:

- Ternary and pseudoternary phase diagrams
- Equations of state (EOS)

Phase behavior (from both methods) provides valuable inputs to the reservoir simulator.

(ii)Advantages of using a compositional simulator

A compositional simulator is the most mechanistically accurate simulator for solvent compositional processes. When the EOS is tuned properly to appropriate experimental data, it computes realistic phase behavior. Thus, the appropriate phase behavior for flooding with enriched hydrocarbon solvent, lean hydrocarbon solvent, N_2, and CO_2 all can be taken into account. Compositional simulators predict the effect of changing pressure and injection-solvent composition on a displacement without the need to enter approximations into the simulator for these effects (except as the EOS itself is an approximation). The compositional simulator is capable of computing realistic behavior when pressure is well below the minimum miscibility pressure (MMP) of the injection solvent, is near but still below the MMP, or is well above the MMP. For this reason, it is ideally suited to study optimum operating conditions.

In addition to these advantages, a compositional simulation, to a large degree, removes the need for a user-defined miscible flood residual oil saturation, as it naturally computes the amount of residual oil left after the interaction of phase behavior and dispersion and distributes this residual saturation realistically as a varying saturation instead of an input, constant saturation.

A compositional simulation can have other aspects of mechanistic reality besides phase behavior. The mechanisms of molecular diffusion and convective dispersion may be included in the equations solved by the simulator. Although grid-refinement sensitivity (described later), or numerical dispersion, may dwarf the effects of these mechanisms in many simulations, they may be important to include in the finely gridded reference simulations (also described later).

Another physical mechanism that can be included in compositional simulations is the effect of interfacial tension (IFT) on solvent/oil relative permeability and capillary pressure. Although one cannot readily foresee the impact of a particular mechanism in the complex compositional simulation of solvent flooding, inclusion of the IFT mechanism seems prudent.When an appropriate relative permeability treatment is included, compositional simulation predicts realistic solvent trapping, especially the trapping of solvent by crossflowing oil. Oil crossflow into a solvent-swept zone immiscibly displaces the solvent in a compositional simulation and leaves the solvent as a residual saturation consistent with the phase behavior.

(iii)Disadvantages of using a compositional simulator

The primary disadvantages of a compositional simulator are the degree of grid refinement often required to compute oil recovery with satisfactory accuracy and the computing time required for fine-grid simulations. These factors generally preclude using a compositional simulator directly for full-field simulations unless some kind of scaling-up technique is used to transfer the information developed from fine-grid reference-model simulations on a limited reservoir scale to coarse-grid simulations on the full-field-model scale. The predicted benefit of compositionally enhanced solvent flooding can be substantially in error if the simulation is made directly with a full-field model with typical coarse grids. This is illustrated by **Fig. 6**, which shows the results of an enriched-solvent-drive reservoir study.[1] In this figure, simulations were made for two one-fourth nine-spot models that represented the same reservoir description.

- One model had a fine grid (30×30×31 cells in the *x*-, *y*-, and *z*- directions)
- The other had the same grid as that used in the full-field model (5×5×17).

The incremental recovery in this figure is the difference between solvent-flood and waterflood simulations in each model. The direct full-field simulation overpredicted incremental recovery by a factor of two.

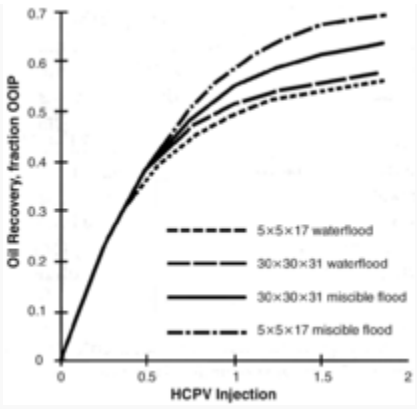

Fig. 6 – Predictions with reference model and corresponding model with full-field grid size

There also are some additional data requirements for predicting solvent trapping and solvent relative permeability hysteresis that are not found in black-oil waterflood simulations.

(iv)Fine-grid reference models

Fine-grid reference models are used to reduce the grid refinement sensitivity problems in compositional simulators.Grid-refinement sensitivity is an extremely troublesome problem in many compositionally enhanced solvent simulations. The problem manifests itself by the predicted behavior changing as the grid is

refined (i.e., as the gridblocks become smaller and smaller). This behavior can be caused by truncation error or numerical dispersion that results from representing derivatives by finite differences; by the inability to accurately resolve the size of solvent tongues or fingers with large gridblocks; and by the inability to represent with large gridblocks some features of reservoir description that have an important effect on solvent sweep, such as discontinuous shales, thin high-permeability strata, or thief zones.

(v)Importance of minimizing grid refinement error

Fig. 7 shows the incremental recovery computed for two different 3D models, one representing one-eighth of a nine-spot pattern, the other representing one-fourth of a nine-spot. Each model had a different geostatistical distribution of correlated permeability with scattered, discontinuous shales represented by zero vertical permeability between gridblocks. Permeability and porosity were scaled up by the renormalization method from the model with the smallest gridblocks to the other models.[2]

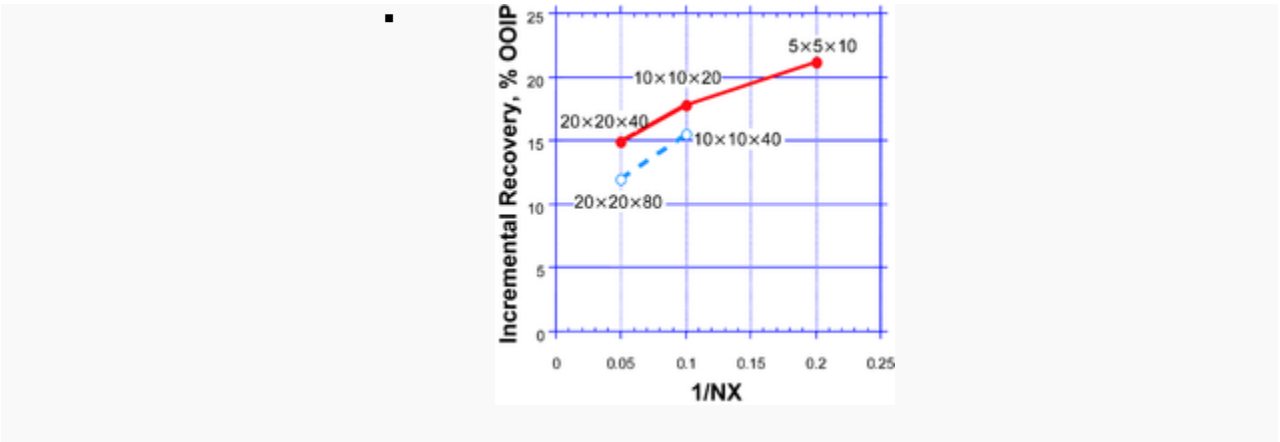

Fig. 7 – Example of grid-refinement sensitivity.

The base model for the one-eighth nine-spot has a grid of 20×20×40. Gridblocks were 93 ft on a side and 1 ft thick. The gridding of the one-fourth nine-spot model was 20×20×80, with gridblocks also 93 ft on a side and 1 ft thick.

Incremental recovery in this figure is plotted vs. 1/NX, where 1/NX is the dimensionless *x*-direction gridblock size. However, in this problem the dimensionless gridblock sizes in the other two directions also vary directly with the *x*-direction gridblock size. It is apparent that as the gridblock size is refined, the predicted incremental recovery decreases for what is supposed to be the same reservoir problem.

Fig. 7 illustrates the importance of minimizing grid-refinement error and explicitly including reservoir-description details that affect flow in an important way. Generally, minimizing the error from grid refinement and accounting for important reservoir-description details adequately requires small gridblocks.

Layers that are 1 ft or no more than a few feet thick and have at least 20 to 40 lateral gridblocks between wells are desirable. Unfortunately, such fine gridding is not feasible for full-field simulations, for most 3D simulations of a single pattern, or perhaps even for some 3D repeating elements of a pattern. Because of this, field predictions need to be made in two steps—with reference models that can be gridded finely enough to accomplish the objectives summarized above, and with scaleup models that incorporate the information derived from reference models into field predictions that account for fieldwide reservoir description, multiple patterns, and operating realities and constraints.

Although it is desirable to make 3D reference-model simulations gridded so finely that the computed answer is adequately close to the converged answer, the discussion above shows that in general, it may not be feasible to do this. A reasonable alternative may be to make finely gridded 2D cross-section simulations to study the grid-refinement issue because for many problems, grid refinement has a larger effect on the computed outcome than the areal effects captured by a coarser-gridded 3D model. Variable-width 2D cross sections sometimes adequately represent the behavior of 3D pattern-segment models with the same fine gridding. In these cross sections, the width is smaller near the injector and producer and increases in the interwell region. This causes flow rate to be greatest near the wells and lowest midway between wells, as it would in a 3D displacement. Even when a fine-grid cross section does not realistically model a fine-grid 3D displacement, it still may predict incremental recovery better than a simulation in a more coarsely gridded 3D model. Moreover, 2D cross-section simulations are well suited for scaleup with the segment and streamline/streamtube models discussed in the next section.

A potential procedure for developing a 3D reference model is first to make a 3D simulation of a pattern element with the finest-grid refinement that is practical. Then, well-to-well cross sections are taken from this model, and the cross sections are refined further. Pseudoproperties are developed for the original cross sections that predict the performance of the more finely gridded cross sections. Then, these pseudoproperties are used in the moderately gridded 3D model to approximate the effect of further grid refinement.[1]

Scaleup to the full field from a fine grid model is the next step in understanding the behavior of a miscible flood.

STIMULATION CONSIDERATION

Many horizontal wells have been completed without plans for stimulation. Often, horizontal wells were not planned for stimulation, because the belief at the time of completion was that horizontal wellbores eliminate the need for hydraulic fracturing stimulation. This presumption has turned out to be false. Often, by the time it is discovered that a horizontal well needs to be stimulated, it cannot be stimulated effectively because of mechanical or reservoir

limitations. This regrettable outcome may have been avoided if extensive preplanning had included consideration for future stimulation. Such preplanning may be limited—not only to the already extensive plans for prospective stimulation activities and selection of the most promising stimulation methods—but for all activities required during the life of the well.

(i)History of Horizontal Wells

Before 1990, US horizontal wells totaled less than 300. By 2004, horizontal wells in the US still numbered less than 4,000, with fewer than 14,000 worldwide (Protecting Our Water 2004; Horizontal and Multilateral 1999). It is likely that most of the horizontal wells drilled before 1990 have depleted to an unsatisfactory production level, now making stimulation a necessity (East et al. 2004), and the US drilling pace for horizontal wells in low-permeability reservoirs has increased since 2005.

Horizontal wells were first drilled in the 1930s, primarily to expose more hydrocarbon-producing rock. Often though, cost and/or risk prevented these types of completions (Ranney 1939). To competitively achieve the same purpose, around 1949, service companies began to offer hydraulic fracturing-stimulation services that proved to be very effective in reaching the unexposed hydrocarbon. This success resulted in a temporary decline of horizontal well technologies, but in the early 1970s, more economical solutions in horizontal well drilling became available. Often during that time, the primary objective was to eliminate the need for costly stimulation and completions.

Eventually, operators began to realize that many of their horizontal wells were not producing as expected. Their options were to abandon the wells, be content with the low production, or stimulate. Usually, hydraulic fracturing stimulation was the desired option, but because the wells had not been completed with future stimulation treatments in mind, fracture stimulations often did not produce satisfactory results. Even when the wells were cased and cemented, many stimulation treatments were marred by screenouts and economically infeasible production increases.

(ii)How Horizontal Wells Differ

Stimulation options in horizontal wells are heavily influenced by the type of completion selected during the design phase. This consideration is less critical with vertical wells, whereby in most cases, any stimulation method can be implemented without unusual pre completion stimulation planning.

A primary difference with horizontal drilling is hydraulic fracture plane position relative to the wellbore. Fig. 8 shows a vertical well (a) that intersects the formation, creating Fractures 1, 2, and 3. Theoretically, no matter what the hydraulic fracture direction, any resulting fracture (i.e., 1, 2, or 3) connects to the wellbore in a similar fashion. That is, a large portion of the fracture

connects the wellbore either axially or longitudinally. Of course, this theory assumes that fractures are always vertical; which, as will be discussed later, may not be the case. After the wellbore is laid down (as in view b), and relative to the wellbore, hydraulic fractures can be positioned in any relative direction imaginable.

To further complicate matters, horizontal wells as defined in the industry, in most cases, are not precisely horizontal. Slants, dips, and "up-and-downs?? are often designed into (or an unintentional result from) the drilling program (a complicated up-and-down horizontal well is shown in Fig. 9). Obviously, most horizontal or deviated wells are not as complicated as the one shown, and often their shape is controlled (and limited) by the capability of the drilling company performing the operation. Another controlling factor is rock characteristics (e.g., the presence of brittle hard rock may incapacitate steering mechanisms in many drilling systems).

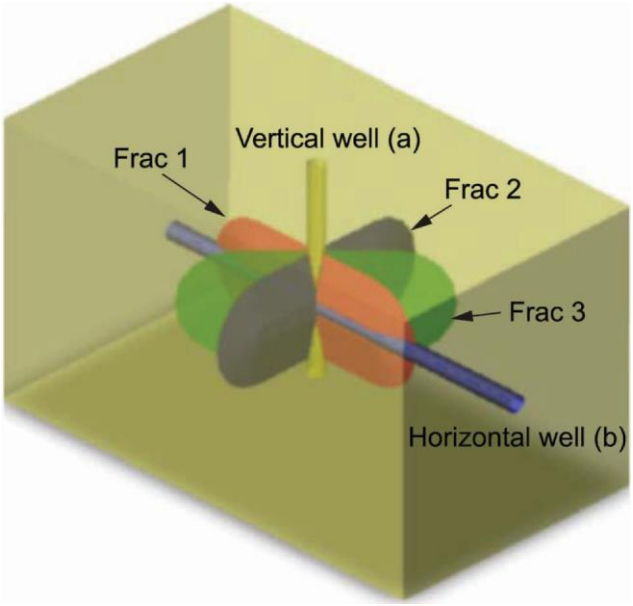

Fig.8 Fractures in a rock formation

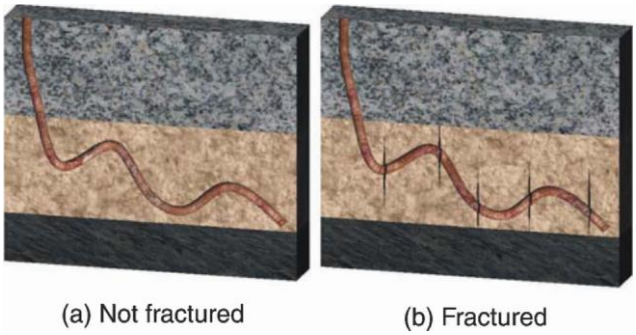

Fig9. Complex structured horizontal wellbore